Enhancing Disaster Situational Awareness Through Scalable Curation of Social Media

DOCTORAL THESIS

Jakob Rogstadius
INFORMATICS ENGINEERING

UNIVERSIDADE da MADEIRA
A Nossa Universidade
www.uma.pt

July | 2014

Enhancing Disaster Situational Awareness Through Scalable Curation of Social Media

DOCTORAL THESIS

Jakob Rogstadius
INFORMATICS ENGINEERING

SUPERVISOR
Evangelos Karapanos

CO-SUPERVISOR
Vassilis Kostakos

Abstract

Enhancing Disaster Situational Awareness Through Scalable Curation of Social Media

by

Jakob Rogstadius

Online social media is today used during humanitarian disasters by victims, responders, journalists and others, to publicly exchange accounts of ongoing events, requests for help, aggregate reports, reflections and commentary. In many cases, incident reports become available on social media before being picked up by traditional information channels, and often include rich evidence such as photos and video recordings.

However, individual messages are sparse in content and message inflow rates can reach hundreds of thousands of items per hour during large scale events. Current information management methods struggle to make sense of this vast body of knowledge, due to limitations in terms of accuracy and scalability of processing, summarization capabilities, organizational acceptance and even basic understanding of users' needs. If solutions to these problems can be found, social media can be mined to offer disaster responders unprecedented levels of situational awareness.

This thesis provides a first comprehensive overview of humanitarian disaster stakeholders and their information needs, against which the utility of the proposed and future information management solutions can be assessed. The research then shows how automated online text-clustering techniques can provide report de-duplication, timely event detection, ranking and summarization of content in rapid social media streams. To identify and filter out reports that correspond to the information needs of specific stakeholders, crowdsourced information extraction is combined with supervised classification techniques to generalize human annotation behaviour and scale up processing capacity several orders of magnitude.

These hybrid processing techniques are implemented in CrisisTracker, a novel software tool, and evaluated through deployment in a large-scale multi-language disaster information management setting. Evaluation shows that the proposed techniques can effectively make social media an accessible complement to currently relied-on information collection methods, which enables disaster analysts to detect and comprehend unfolding events more quickly, deeply and with greater coverage.

Keywords: *social computing, crisis mapping, disaster response, text mining, crowdsourcing, social media.*

Resumo

Enhancing Disaster Situational Awareness Through Scalable Curation of Social Media

por

Jakob Rogstadius

Actualmente, mídias sociais são utilizadas em crises humanitárias por vítimas, apoios de emergência, jornalistas e outros, para partilhar publicamente eventos, pedidos ajuda, relatórios, reflexões e comentários. Frequentemente, relatórios de incidentes estão disponíveis nestes serviço muito antes de estarem disponíveis nos canais de informação comuns e incluem recursos adicionais, tais como fotografia e video.

No entanto, mensagens individuais são escassas em conteúdo e o fluxo destas pode chegar aos milhares de unidades por hora durante grandes eventos. Actualmente, sistemas de gestão de informação são ineficientes, em grande parte devido a limitações em termos de rigor e escalabilidade de processamento, sintetização, aceitação organizacional ou simplesmente falta de compreensão das necessidades dos utilizadores. Se existissem soluções eficientes para extrair informação de mídias sociais em tempos de crise, apoios de emergência teriam acesso a informação rigorosa, resultando em respostas mais eficientes.

Esta tese contém a primeira lista exaustiva de parte interessada em ajuda humanitária e suas necessidades de informação, válida para a utilização do sistema proposto e futuras soluções. A investigação nesta tese demonstra que sistemas de aglomeração de texto automático podem remover redundância de termos; detectar eventos; ordenar por relevância e sintetizar conteúdo dinâmico de mídias sociais. Para identificar e filtrar relatórios relevantes para diversos parte interessada, algoritmos de inteligência artificial são utilizados para generalizar anotações criadas por utilizadores e automatizar consideravelmente o processamento.

Esta solução inovadora, CrisisTracker, foi testada em situações de grande escala, em diversas línguas, para gestão de informação em casos de crise humanitária. Os resultados demonstram que os métodos propostos podem efectivamente tornar a informação de mídias sociais acessível e complementam os métodos actuais utilizados para gestão de informação por analistas de crises, para detectar e compreender eventos eficientemente, com maior detalhe e cobertura.

Palavras chave: *computação social, mapeamento de crises, resposta humanitária a crises, extração de texto, crowdsourcing, mídias sociais.*

Acknowledgements

Sincere thanks go to Syria Tracker and other study participants for their feedback on the practical utility of the developed systems, and to Ko-Hsun Huang for her methodological guidance and support throughout the research. The research was co-supervised by Evangelos Karapanos at the University of Madeira and Vassilis Kostakos at the University of Oulu.

The work was supported in part by an IBM Open Collaboration Award; by the Portuguese Foundation for Science and Technology (FCT) grant CMU-PT/SE/0028/2008 (Web Security and Privacy); by NSF grants OCI-0943148 and IIS-0968484; and through an internship in the Social Computing group at QCRI.

Publications

This doctoral thesis includes research previously presented in the following publications, in addition to previously unpublished work:

i. Rogstadius, J., Kostakos, V., Laredo, J. and Vukovic, M. (2011). "Towards Real-time Emergency Response using Crowd Supported Analysis of Social Media". In: *CHI 2011 Workshop on Crowdsourcing and Human Computation.* Vancouver, Canada.

ii. Rogstadius, J., Kostakos, V., Kittur, A., Smus, B., Laredo, J. and Vukovic, M. (2011). "An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets". In: *Proc. ICWSM11.* Barcelona, Spain. pp. 321-328.

iii. Rogstadius, J., Kostakos, V., Laredo, J. and Vukovic, M. (2011). "A real-time social media aggregation tool: Reflections from five large-scale events". In: *ECSCW 2011 Workshop on Collective Intelligence and CSCW in Crisis Situations.* Aarhus, Denmark.

iv. Rogstadius, J., Teixeira, C., Karapanos, E., and Kostakos, V. (2013). "An Introduction for System Developers to Volunteer Roles in Crisis Response and Recovery". In: *Proc. ISCRAM13.* Baden-Baden, Germany.

v. Rogstadius, J., Teixeira, C., Vukovic, M., and Kostakos, V., Karapanos, E., Laredo, J. (2013). "CrisisTracker: Crowdsourced Social Media Curation for Disaster Awareness". In: *IBM Journal of Research and Development* 57.5, pp. 4:1-4:13.

vi. Imran, M., Castillo, C., Lucas, J., Meier, P., Rogstadius, J. (2014). "Coordinating Human and Machine Intelligence to Classify Microblog Communications in Crises". In: *Proc. ISCRAM14.* State College, PA.

The thesis author's contributions to these publications are listed in Table 1, together with the thesis chapters in which material from the publications is included.

	Concept or software design	Software development	Study design	Study execution and analysis	Literature review	Report writing	Thesis chapter
i	◐	n/a	n/a	n/a	●	◐	1,2,8
ii	◐	●	◐	●	○	◐	4
iii	●	●	●	●	●	●	5
iv	◐	n/a	○	○	◐	●	1,2
v	●	◐	●	●	◐	●	3,6
vi	◐	◐	○	○	●	◐	7

TABLE 1: The thesis author's contributions to publications contributing to this thesis. ● Primarily thesis author; ◐ collaborative work; ○ primarily co-authors.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

In the period of time following a natural disaster or other large scale emergency, regular citizens have long suffered from poor situational awareness. An individual's information access has largely been restricted to direct observations of their immediate surroundings, combined with sparse high level summaries disseminated through mass media. However, access to information is as critical to disaster affected populations as is water, food and shelter (IFRC 2005).

In recent years, online social media and other emerging ICTs have enabled regular citizens to participate actively in information exchange during large-scale events, such as earthquakes, elections, storms, bushfires and terrorist attacks. Social media is now used to share timely first hand information about hazards, interventions, fatalities, personal status and damage, in addition to the summary reports, reflections and commentary that would traditionally be associated with journalistic news media (Vieweg 2012). Furthermore, reports of new incidents and new developments of ongoing incidents can get picked up earlier and sometimes with greater coverage in social media than by traditional media channels such as TV, radio and news websites (Rogstadius et al. 2013b). Simultaneously, the pervasiveness and robustness of infrastructure required to participate in the online information exchange is improving rapidly around the globe, and distributed real-time information exchange between disaster affected citizens is becoming the new norm.

Timely citizen reports posted on social media can bring several positive changes to the disaster landscape. Beyond the primary effect of direct communication and coordination between citizens, this new media has the unique property that much of the communication can be accessed publicly by third parties. Large-scale mining of social media, at least in theory, introduces the possibility of tapping into the collective awareness of a disaster affected population, to leverage thousands or even millions of people as a distributed sensor network. Furthermore, centralized

information collection conducted by multiple virtually coordinated entities has potential to minimize redundant relief efforts and speed up initial disaster response (Verity 2011). This creates opportunities for unprecedented levels of situational awareness among emergency managers, trained responders and other decision makers in humanitarian disasters.

Several technologies have been proposed to make information in public online communications more accessible to decision makers. These approaches rely either on automated or human computation to collect, organize, and visualize information to end users. In practice, computational information extraction techniques developed for well-formatted documents have so far fallen short when applied to disaster-relevant social media content. As a consequence, there has been a resurgence of fully manual information processing to achieve results of acceptable quality, made possible through the application of crowdsourcing techniques, which distribute the great manual workload over many individuals. Human computation on the other hand comes with its own set of problems, primarily relating to scalability, speed and cost.

1.2 Problem statement

Research into scalable computation techniques applicable to large-scale mining of social media content during ongoing disasters is a young but rapidly expanding field, with a great number of challenges remaining to be solved. The following problems are considered in this research.

Humanitarian disasters used to be environments characterized by information scarcity. Instead, with the introduction of public online citizen communication, the new standard has become information overload. It is currently not uncommon to see hundreds of thousands of messages being authored every hour during events affecting millions of people. Although timely and informative reports are in theory accessible, the majority of published messages contain personal commentary or information that is otherwise irrelevant to the response effort. Current information management tools are not sufficiently capable of separating signal from noise in this unstructured torrent of text (Gao, Barbier, and Goolsby 2011). As a consequence, analysts and others struggle to make use of potentially life-saving information published through this novel channel. **Minimizing information overload while maximizing situational awareness is the central fundamental issue that currently needs to be addressed by any research in this domain**.

Situational awareness (SA) is a state of understanding a situation as a whole; knowing what is going on around you in relation to the goals and decisions that need to be made (Endsley 2000). Thus information that contributes to SA needs to give a representative view of the disaster environment, summarizing key events and conditions that are relevant to the task at hand.

The challenge of making sense of large document collections is not a new one, and an extensive body of research has been produced in fields such as natural language processing, information extraction, search, machine learning, document clustering, text summarization and visualization. However, this research has almost exclusively focused on document collections such as news articles, books, scholarly articles and web pages. These documents differ from social media content in several significant ways (see section 3.1), such as length, formality of language, and availability of meta-data. Together these differences have significant negative impact on the quality of output of established text processing algorithms.

Furthermore, it has been observed that the information contained in individual short messages often does not contain sufficient detail for planning relief efforts (Morrow et al. 2011). However, as it is not unusual to find multiple independent accounts of the same event, more sophisticated techniques are needed that can **aggregate information from several related reports into comprehensive stories describing unfolding events**. At the same time, access to individual source reports (e.g. SMS or public social media posts) needs to be maintained (Gao, Barbier, and Goolsby 2011), for instance for end users to be able to find the most credible among multiple conflicting statements.

Report redundancy is in fact a significant challenge of its own, as many online social media platforms use message duplication as a method to propagate information between users. This results in high numbers of reports containing equivalent information during large-scale events, which effectively hides less reported events. New techniques are thus needed to better **handle detection and removal of duplicate information** (Gao, Barbier, and Goolsby 2011).

Due to the current shortcomings of systems built using fully automated techniques, volunteer-based crowdsourced human computation has become the de-facto standard for social media monitoring during disasters. Humans are a flexible resource, easily adapted to seek out new information or meta-data types, to process reports in new formats or languages, to process images and videos, or to disambiguate and interpret short contextual messages.

However, human computation is highly labour-intensive, and individual workers are prone to burn-out, which makes volunteer-based human computation unsuited for prolonged disasters, e.g. civil wars. In addition, human computation is slow compared to automated processing. Out of tens of thousands of reports submitted to a landmark deployment of the most popular crowd-reporting platform, only around five percent were sufficiently processed to become accessible to system users (Morrow et al. 2011). What's worse, since the system had no built-in method for prioritizing which reports to process first, urgent messages may not have been prioritized, and responders who accessed the curated data questioned the representativeness of the sample. Therefore, research is needed to find **scalable approaches for extracting meta-data**, which is required to index reports and then match it with relevant information consumers.

In addition to ranking content for human computation, solutions are needed for **timely breaking news detection**. In a constantly flowing stream of information, which is far too rapid for any one person to process, a system should be able to extract information pieces that are particularly important to decision makers, while keeping the rate at which new information pieces are highlighted low enough to not cause information overload. Furthermore, to offer advantages over existing practices, the process must be able to surface breaking news with greater timeliness and/or reliability than other channels already in use, such as TV, online news websites or field-deployed staff.

A possible approach to increase the scalability and sustainability of human computation is to use crowdsourcing markets such as Amazon's Mechanical Turk. However, the workers of these markets seek financial reimbursement for their time, which may or may not be available in disaster settings. Research is needed to **investigate if workers on task markets are willing to work for less payment if the work contributes to social good**.

An in-depth understanding of end-user needs is critical to the success of any information management solution. However, despite the documented value of some of the new platforms, it remains necessary to more clearly define and document various levels of decision makers' needs in terms of information and products during the various phases of a humanitarian crisis (Verity 2011). For instance, a study assessing the real value of an extensive human effort to classify reports into topic categories established that the categories used corresponded poorly with data structures and needs of formal organizations (Morrow et al. 2011). **A general assessment of decision maker's needs** is therefore needed to develop meta-data structures and classification schemes for reports that meet the needs of end users. This is true regardless of if the classification is performed by a human or a computer.

Finally, proposed methods and tools should be evaluated under realistic settings, as closely as possible corresponding to those of real-world disaster information management. This is to **verify that information extracted is of high utility at the time when the system presents it to a user**. As social media is only one among many information sources available to end users, it is necessary to understand not only what information a system can provide, but also how this information compares in richness, timeliness and credibility to information received through other information collection methods already in use.

To summarize, several problems need to be solved before online social media can become useful as an information source in humanitarian disaster response. First, the information needs of decision makers in different stakeholder groups need to be better understood, so that software tools can be designed to provide information in a format that improves situational awareness. Scalable, accurate and timely methods then need to be identified to handle processing of very rapid streams of short messages, including novelty detection, de-duplication, summarization,

meta-data extraction and presentation. Finally, solutions should be evaluated under as realistic settings as possible.

1.3 Summary of contributions

The primary audience of this thesis is researchers and system developers working in crisis computing and disaster information management. The research has the following key contributions, discussed in greater detail in chapter 8:

- Specific information types are identified that improve the situational awareness of key stakeholders in response to humanitarian disasters (chapter 2), followed by an assessment of available methodologies for collecting and managing citizen-generated information (chapter 3).
- An assessment of the effects of intrinsic and extrinsic motivation on work posted on crowdsourcing markets (chapter 4). The study suggests that humanitarian value alone is not sufficient to enlist workers on for-pay markets, and that volunteer-based human computation is likely to generate higher quality output in most disaster-related applications.
- With current technologies, effective social media monitoring requires hybrid workflows, in which human-based computation is made scalable through support by machine-based techniques. Hydbrid processing is implemented in the open-source CrisisTracker system, which through field evaluation is demonstrated to effectively improve the situational awareness of international disaster analysts (chapters 6-7).
- A demonstration of how automated online text-clustering techniques can be used to provide report de-duplication, timely event detection, ranking and summarization of content in rapid social media streams (chapters 5-6).
- A technique is proposed for scalable meta-data extraction for detected events, using hybrid processing in which supervised classifiers learn to generalize human tagging behaviour to message volumes several orders of magnitude greater than what can be processed using human-based computation alone (chapter 7). Meta-data extraction serves to identify and filter out reports that correspond to the information needs of specific stakeholders.

1.4 Delimitations and scope

This research is solely focused on supporting the response phase of disaster management, i.e. the time during and immediately after a disaster when the disruption to society is greatest. The

research is also specifically focused on disasters, which affect large parts of a community or even an entire country, as opposed to smaller-scale emergencies that are well-bounded in time and space, such as a property fire, a plane crash or limited riots.

Furthermore, the work aims to find information management approaches suitable for use in sufficiently developed but disaster-prone regions of the world. Intended use-cases include monitoring the impact of natural disasters, civil unrest and conflict, that affect regions in the upper and middle segments of the economic development spectrum. The least developed regions of the world have yet to experience a boom in internet connectivity, and thus lack sufficient communications infrastructure for social media monitoring to be a valid method of information collection.

Evaluation of the research has only been conducted in settings representative of the center of the economic development spectrum. In the observed settings, humanitarian organizations have had greatly limited resources at their disposal. Therefore, cost, both in money and manpower, has been a significant concern in information collection and analysis. Further research may be necessary before generalizing the findings of this research to other contexts.

The work presented herein also focuses exclusively on increasing the utility of a single information channel; online social media. Disaster response is complex, and in practice this channel is merely one of several from which responders collect and aggregate information. This means that the systems developed during this research are also only useful during disasters in communities that actively use social media. The value of the work is also reduced in disasters that severely impact communications infrastructure for the public for lengthy periods of time.

Chapter 2

The value of monitoring citizen communication during disaster response

This chapter discusses why and how citizen communication on online social media during disasters has the potential to give disaster response organizations unprecedented situational awareness. After defining the concepts of humanitarian disasters and situational awareness, the chapter looks back at how regular citizens in a disaster environment have tried to satisfy their information needs historically and why social media is such a disruptive technology. The chapter then looks at disaster situational awareness in practice, identifying which stakeholders need what information in their decision making. Finally, conclusions are drawn that can inform design and feature selection of software tools for information management in disaster response.

2.1 Humanitarian disasters

The International Federation of Red Cross and Red Crescent Societies (IFRC) defines a **disaster** as a sudden, calamitous event that seriously disrupts the functioning of a community or society and causes human, material, and economic or environmental losses that exceed the community's or society's ability to cope using its own resources. Though often caused by nature, disasters can have human origins (IFRC 2013b).

A **humanitarian disaster** (or 'humanitarian crisis') is an event or series of events which represents a critical threat to the health, safety, security or well-being of a community or other large group of people, usually over a wide area. Armed conflicts, epidemics, famine, natural disasters and other major emergencies may all involve or lead to a humanitarian crisis (Humanitarian Coalition 2013).

To reduce the impact of disasters (and smaller emergencies), local, national and international emergency management agencies have been put in place. These specialize in **emergency management**, which is commonly described in a four-phase model consisting of mitigation, preparedness, response and recovery (Queensland Government 2013). Mitigation deals with taking precautionary steps to avoid that an emergency happens in the first place, by increasing resiliency to extreme events. This includes building flood protection systems, preventing construction in emergency prone areas, and in other ways improving infrastructure. Preparedness deals with planning and training for the event that an emergency still happens. Response is the immediate actions following an emergency situation and recovery is the long-term process of returning a community to the pre-emergency state.

Disasters can also be classified as either rapid onset or slow onset (IFRC 2013a). Rapid onset disasters such as earthquakes, tsunamis, storms, or flash floods, are typically characterized by a short event causing wide-spread property damage, casualties and disruption to basic utilities and services, followed by a longer period of community recovery efforts. In slow onset disasters, such as drought, disease outbreaks or civil war, it is more difficult to define where the response phase ends and recovery begins, as threats to the community remain for a much longer period of time, only gradually disappear, and may resurface in new locations. In both rapid and slow onset disasters, signals can be available at an early stage that permits deployment of preventive and reactionary measures.

2.2 Situational awareness

Victims, citizens in disaster-affected communities, members of formal response agencies, and concerned outsiders all gather available information before deciding what action to take regarding an emergency. This process of gathering information, or situational assessment, leads to a state of situational awareness (Vieweg 2012).

Situational awareness (SA) is an elusive concept that many authors have attempted to define. An extensive review by Nofi (2000) finally settled on the definition given by Endsley (2000), which will be used throughout this thesis. Endsley states that *"situational awareness is a state of understanding a situation as a whole; knowing what is going on around you in relation to the goals and decisions that need to be made."* This is formalized into four processes:

Perception Acquisition of available facts;

Comprehension Understanding the collected facts in relation to goals and objectives;

Projection Envisioning how the situation will develop without outside influence;

Prediction Envisioning how outside forces will affect the projection.

What constitutes SA differs by an individual's role and situation. For instance, the things that are relevant to a member of a search-and-rescue team during the immediate aftermath of an earthquake are not the same as those relevant to a remotely located information management officer in an international organization during the same event. Similarly, the information needed during the mitigation phase of disaster management is different from that needed during the response phase.

Furthermore, SA changes continuously with an evolving situation; it is an ongoing process, or a working hypothesis. The four processes in Endsley's model take place in parallel, rather than sequentially (Nofi 2000). However, Endsley does organize the four processes into a three-level dependency model, with projection and prediction first requiring comprehension, and comprehension first requiring perception. Comprehension, projection and prediction are all strongly facilitated by training and experience, which gives an operator adequate mental models in which to organize observations.

2.3 History of citizen communication in disasters

Past decades' advances in information and communication technologies (ICTs) have substantially changed how information gets disseminated during large-scale complex events, such as natural disasters, conflict and political unrest. Physically bounded means of communication, such as handing out fliers with pictures of missing persons and putting up posters with requests and offers (Palen and Liu 2007) have gradually been replaced or supplemented by technology-mediated communication.

Traditional mass media, in particular television, is capable of transferring powerful live footage across national borders. Its contribution to mass movements that led to several regime changes has been documented, including helping to topple the Berlin Wall in 1989 (Lambert 2005), inspiring widespread protests against Suharto in Indonesia in 1998 (Lambert 2005), and raising local and international awareness of the scale of citizen uprisings in Libya and Egypt in 2011 (Harb 2011), despite government efforts to tone down the events. During the Tiananmen Square protests in China in 1989, government authorities jammed international radio and television to isolate the protesters. Instead, fax machines were used to get the word out to the world and to communicate international support back to protesters (Lambert 2005; Houle 2009).

The introduction of the cell phone enabled citizens to directly communicate with each other, without being physically confined to the home or work place. Text messages (SMS) helped organize and overthrow the rule of Joseph Estreda in the Philippines in 2001 (Lambert 2005), and were used extensively for peer-to-peer communication during the 2003 SARS epidemic in China (Palen and Liu 2007). After the devastating earthquake in Haiti in 2009, SMS coverage

was quickly restored and enabled victims in destroyed areas to communicate information and needs to aid workers (Harvard Humanitarian Initiative 2011). Camera phones with MMS were also used during the 2005 London Tube Bombings to transmit citizen footage to news media and authorities (Palen and Liu 2007).

The personal computer and the Internet further empowered citizens with better information discovery, many-to-many communication and improved ability to discover likeminded peers. For instance, online newsgroups and email were used to coordinate largely leaderless protests in Indonesia in 1998 (Lambert 2005) and people used library computers to search for information, locate missing family members and to communicate with relatives and friends in the aftermath of hurricanes striking the US Gulf coast in 2004 and 2005 (Jaeger et al. 2006). Following Hurricane Katrina in 2005, institutions normally responsible for providing relief were overwhelmed and affected citizens and donors organized themselves in blog communities and on discussion forums to connect needs with donated resources (Torrey et al. 2006). Citizens also used online discussion forums to exchange information and coordinate after the 2008 Sichuan earthquake in China, and a web platform was set up to aggregate hospital records, citizen reports and other information to help connect missing people with relatives (Qu, Wu, and Wang 2009).

2.3.1 Social media

With widespread adoption of social media, together with handheld devices with internet connectivity, GPS and cameras, the information space during large-scale complex events has become ever more connected. This has given rise to citizen journalism, for instance during the 2008 Mumbai Attacks when eye witnesses were able to document their experiences in real-time by uploading photos and sharing live updates through services like Twitter and personal blogs (Stelter and Cohen 2008).

By the end of 2012, the number of mobile-connected devices was expected to exceed the world's population (Cisco 2012) and 34% globally had access to the Internet, with connectivity rising rapidly in Africa and the Middle East (Internet World Stats 2012). By the end of 2011, Facebook had close to 1 billion users and 100 million active Twitter users sent over 1 billion tweets every week (Turrell 2011), numbers that grow every month. Despite that broadband adoption in the Middle East is only 1-7%, 24 hours of video is being uploaded to YouTube every minute (Ghannam 2011) and during protests in 2011 Egyptians used Facebook, Twitter, and YouTube to post millions of internet links, news, articles and videos to people all over the world (Bhuiyan 2011).

New infrastructure, devices and media have removed the time-lag between an event and the delivery of news about that event that was built-in to traditional media (Faris 2008). Furthermore, the dramatic reduction in communication cost and real-time nature of social media facilitates

both organization and collective action (Meier 2011). People no longer only seek response- and rescue-relevant data, but opportunistically and actively provide it as well, including information about structural damage, flooding, places where people need to be rescued, and missing person searches. They also seek and provide relief assistance, for instance information about housing, food, jobs and transportation help (Palen and Liu 2007). Content analysis of tweets collected during natural disasters indicates great availability of response-phase related reports: hazards, interventions, fatalities, personal status and damage (Vieweg 2012). Usage even goes beyond basic reporting and also includes aggregate reports, reflections and commentary that would traditionally have been associated with traditional news media (Faris 2008). An Egyptian activist in 2011 explained the use of new communication technologies as *"We use Facebook to schedule the protests, Twitter to coordinate, and YouTube to tell the world"* (Howard 2011).

Social media is used throughout the emergency management cycle to detect potential hazards, gain situational awareness, engage and mobilize local and government organizations and to engage volunteers at the disaster recovery stage. Users of social media at disaster time include victims, volunteers, and relief agencies.

A critical aspect of these new communication channels is that some online social networks, in particular Twitter, provide public API access to real-time feeds of data. While the data can suffer from strong sample bias on a local scale, larger-scale patterns can accurately match those obtained through traditional data collection methods (Graham, Poorhuis, and Zook 2012; Graham 2012). In theory, this makes it possible for emergency managers, community leaders and other decision makers to access live reports generated by a vast distributed network of people within seconds or minutes of events taking place, almost anywhere in the world. In practice, this becomes a challenging filtering problem, as the issue is no longer information scarcity but information overload. The high resolution and rate at which information is being generated during mass events means that decision makers need access to powerful tools that help extract actionable reports and reveal interesting patterns if they wish to tap into the collective situation awareness of the crowd.

2.3.2 Twitter

This thesis focuses primarily on use of information from Twitter, and it is therefore worth describing this specific service in more detail. At its core, Twitter[1] is a microblogging service that enables its users to send and receive short messages, called tweets, of up to 140 characters. It is increasingly used for information sharing (Hughes and Palen 2009) and has emerged as one of the main tools for real-time information exchange between geographically distributed

[1] http://twitter.com/

individuals. As of September 2013, Twitter has over 230 million active users (Twitter 2013) who use the service for socializing, activism, and other non-crisis activities (Java et al. 2007).

Twitter's use has been documented during hurricanes (Hughes and Palen 2009), wildfires (Sutton, Palen, and Shklovski 2008), earthquakes (Harvard Humanitarian Initiative 2011), school shootings (Vieweg et al. 2008) and political protests and conflict (Harb 2011; Bhuiyan 2011; Meier 2011; Howard 2011; Faris 2008). As previously mentioned, there is a documented prevalence of response-phase related reports on the service, including mentions of hazards, interventions, fatalities, personal status and damage (Vieweg 2012), and some users provide additional aggregate reports, reflections and commentary (Faris 2008).

Critical to the work presented here, Twitter differs from many other communication platforms in that most communication is public in nature, and messages can be accessed through text search both by anyone using the service and through API endpoints.

During large-scale events affecting densely populated areas, activity on Twitter can reach hundreds of thousands or even millions of on-topic messages per hour (Chowdhury 2011; Konkel 2013). However, this number should be taken with a grain of salt. Much of this content is redundant, as duplication is a primary driver of how information spreads in the platform. By design, users are encouraged to re-share verbatim copies of messages authored by others (called re-tweeting), or at other times re-word one or several messages into new posts. In addition, as this research will show, many users independently report the same information following similar experiences during mass events. The extent of this information duplication and techniques for de-duplication will be investigated in later chapters.

2.4 Stakeholders and information needs

The disaster environment is highly complex, with presence of a large number of stakeholder groups. This section aims to give an overview of these stakeholders, as well as their basic role in an affected community's response to disaster. The specific information needs of a subset of the stakeholders will also be discussed. The goal is not to provide an exhaustive description, but rather to explain the general setting in which this thesis aims to make a contribution by introducing novel information management techniques.

Since roughly around the same time as work began on this thesis, there has been a growing interest in clear documentation of information needs of the different stakeholders in disasters. At a high level, the categorization of stakeholder groups in this chapter is based on the humanitarian decision maker map by Verity (2013). The description of each stakeholder group's role and information needs is based on material assembled from academic literature, practitioner

manuals, field studies, interviews, and private discussions. To the author's knowledge, this section forms one of the first attempts at a more comprehensive and comparative review of the information needs of different stakeholder groups.

Much of the original source material for this section was published first after the research in the later chapters was conducted, thus some discrepancies do exist between the highlighted needs and the ideas presented in later chapters. These discrepancies will be addressed in detail in the final discussion chapter, many of which directly point to directions for future work.

Unless stated otherwise, sections 2.4.1 and 2.4.2 are based on original research first published in (Rogstadius et al. 2013a). Readers interested in developing information management tools for these stakeholder groups are directed to this paper, which includes two case studies of public sector organizations and volunteers during community response to natural disasters, as well as implications for system development.

2.4.1 Victims and on-site volunteers

The largest and most diverse stakeholder group in any humanitarian disaster is civilian individuals. These include injured or displaced community members; community members who volunteer or consider volunteering in collective or individual response or recovery; non-residents who arrive at the scene with an intent to help; and bordering communities that see an influx of displaced persons (Rogstadius et al. 2013a).

Emergency management literature often describes affected individuals only as passive or incapacitated victims, or sometimes as human resources that can be incorporated into emergency management organizations (Dynes 1994). Unsolicited involvement in emergency response activities has been documented to be disruptive to emergency response organizations, and is therefore generally discouraged (Green III 2003; Lowe and Fothergill 2003).

However, in a humanitarian disaster, the scale of destruction and resulting community needs far surpasses what the community's regular response organizations can handle. In disasters, successful response and recovery depends to a great extent on the collective efforts of large numbers of spontaneous volunteers, who take over tasks that are under more normal circumstances handled by officials. In particular, volunteers make up the first line of response, during the hours, days or even weeks before trained local, national or international response teams arrive at the scene (Brennan, Barnett, and Flint 2005; Quarantelli 2008; United Nations Volunteers 2011).

Access to credible information represents a basic, urgent need for disaster affected populations, as much as food, water, healthcare and shelter (IFRC 2005). The information needs of individual members of the affected community were investigated in and include knowledge of hazards and damage levels in the individual's neighbourhood or local surroundings. In addition, affected

individuals seek out information that helps them assess when help will arrive or basic utilities will be restored. If the arrival of help is unlikely, the individual may look for shelter locations and evacuation routes, or seek out self-help advice and documented best practices. Primarily in larger disasters, there is also a great need for information regarding the status of family members and friends.

Many community members will also want to know what opportunities they have to help others more severely affected than them. Volunteering can be triggered by a combination of several factors, such as personal lack of injury; awareness of nearby needs; low likelihood of trained responders providing a timely solution; and awareness of fellow community members or trained responders with whom to collaborate. Such "spontaneous volunteers" tend to address problems that pose limited personal risk, such as clearing debris and obstacles, providing food, or offering psychological support (Brennan, Barnett, and Flint 2005; Dynes, Quarantelli, and Wenger 1990; Gonzalez 2005; Lowe and Fothergill 2003; Stallings and Quarantelli 1985). The more organized the volunteers become, the more their information needs will resemble those of public-sector organizations.

Individuals outside of the community have less acute information needs, but are commonly seeking general information about the disaster, the status of potentially affected friends and relatives, and concrete ways in which they can help.

As information seekers, members of the public lack the communication infrastructure and organizational protocols now prevalent among professional responders (Vieweg 2012). Research thus shows that in times of crisis, people turn to social media like Twitter to access and broadcast information that can potentially contribute to situational awareness (Vieweg et al. 2010).

Members of a disaster affected community have unique first-hand awareness of local pre- and post-disaster conditions, and can become valuable information producers. They can do this by reporting their observations via SMS, phone calls, e-mail, social media, blogging and other means of communication. Reports matching the information needs of public sector and international organizations are all valuable for the collective response community, including information on infrastructure damage, personal injuries, disease, material needs, hazards, population movements and general "events".

Purpose-built reporting systems and organizational practices to use citizen reports are an active area of research and development, and the information producer role of volunteers is discussed further in section 2.3 later in this chapter.

2.4.2 Public sector organizations

The public sector of an affected community includes emergency management organizations such as fire fighters, law enforcement, search and rescue teams, sea rescue, and medical services; government; intelligence services; and utility companies providing water, electricity, gas, petrol and telecommunications. Though some of these may technically be privatized, the term public sector will be used here for simplicity.

The collective preparedness of the public sector to a great extent determines the resiliency of the community. This includes preventive measures such as restricting construction in disaster prone areas, developing disaster management plans, training, and executing those plans once disaster strikes.

Dedicated emergency management organizations are the community's default approach to handle day-to-day emergencies, and can be assumed to exist in some form before a disaster strikes. Their extensive training combined with high local awareness makes them uniquely skilled to handle challenges effectively and efficiently. While organisational structures vary by country, emergency management organizations are typically separated into national, regional and local levels. With increasing scales and severity of emergencies, organizations first deploy individual operational teams, then local organizations, then regional, national and eventually international resources. Resources of many types can be deployed, including equipment, operational teams, operational management, and capacity for information collection and management.

The emergency response resources available to public sector institutions vary greatly around the world. In highly developed societies, response organizations have far greater access to training, equipment, communication infrastructure and staff. Less developed, or less institutionalized, societies will more often need to rely on citizen-based response to everyday crises.

However, during a true humanitarian disaster, the public sector organizations are greatly overwhelmed by the scale of description and resulting community needs, forcing strict prioritization of resources towards solving the most critical problems at hand (Dynes, Quarantelli, and Wenger 1990; Quarantelli 2008; Rogstadius et al. 2013a). To achieve greater organizational scalability and overall resilience, some countries therefore have extensive volunteer programmes in which civilians can register pre-disaster and receive training, to be incorporated as additional manpower when disaster strikes (Emergency Management Australia 2006).

During disasters, response organizations have also been observed to face new challenges in terms of interfacing and coordinating with several new entities that are not active during regular emergency response activities. These entities can include international response teams, spontaneous volunteers and independent online information management systems. As new communication technologies greatly empower regular citizens with communication and information discovery,

there are important implications for formal organizations. A recent survey by the American Red Cross shows that the vast majority of citizens surveyed believe that national response organizations should regularly monitor social media sites in order to respond promptly. In fact, more than one-third of those polled said that they would expect help to arrive in less than an hour after posting a need online during a crisis (Meier 2011). While this may never be feasible in truly large-scale disasters simply due to resource constraints, tools are needed that help organizations reliably detect new events in minutes rather than hours, so that responders can prioritize where to allocate their scarce resources. Traditional crisis response organizations built around the command-and-control model will also need to find ways to adapt to increasingly common improvised activities by the public (Palen and Liu 2007; Quarantelli 2008).

The information needs of the public sector organizations include knowledge of victims' status and location; infrastructure damage; existing or anticipated hazards; implications for logistics; reports of new events; weather forecasts; communication channels to other responders; the activities of other responders; city plans, infrastructure plans and other geographical data; and population registers. In addition, the large and mostly hierarchical organizations maintain extensive information records regarding their own operations, such as the location and activity of personnel, vehicles and other equipment, typically using specialized software for resource and inventory management.

Naturally, information needs change with roles and responsibilities up and down the organizational hierarchy. At operational level, a team such as a search-and-rescue or fire fighting unit require detailed awareness of everything relating to their immediate task, while having less need for high-level aggregate summaries of the disaster as a whole. Local, national and international operational teams are therefore backed up by some form of information management capacity, which quickly can put together relevant briefings on landmarks, risks, victims, resources and more pending a new deployment (Harvard Humanitarian Initiative 2011).

Further up the chain of command, at local tactical level, an incident commander or equivalent is responsible for directing operational resources towards handling the highest priority needs of the community. Primarily, this requires awareness of needs, resources at disposal, and capacity to anticipate how the situation will unfold.

Most information management takes place at strategic level. Here reports from the different operational branches are assembled and fused with other datasets, such as SOS call centre reports, simulations of fire, smoke, weather and flooding, topological maps, satellite photography, social media reports, and news footage, into some form of continuously updated model of the disaster environment. Both past, present and future events are tracked. Tactical and operational levels can request information as needed.

In truly large-scale disasters, international assistance may be necessary and more organizations need to communicate with each other, but the basic information needs remain similar.

Branches of government that are not specific to emergency management, such as meteorological offices, social services and road maintenance departments, still play a significant role in community recovery. They support response organizations with additional resources, supervise the relief efforts, and in disasters some can serve as authoritative coordinators to connect external donors with local distributors.

The bulk of information flowing from formal response organizations to the affected public typically consists of summaries of ongoing efforts by the organizations, summaries of affected areas, evacuation orders, and other reporting following a mass-media pattern. In the light of the information needs of affected populations described in the previous section, it is perhaps not surprising that affected individuals have been observed to express frustration with the one-way nature of this communication, as well as to approach operational staff on the ground for additional personally relevant information, with varying success. To the author's knowledge, no services have ever been set up by professional response organizations to enable individuals to tap into the detailed situational models maintained by professional information managers, to access relevant information at a hyper-local level.

2.4.3 International organizations

In the most impactful humanitarian disasters, the resources within the affected region are greatly overwhelmed. Frequent effects of disasters include damaged critical infrastructure, disruptions to regular response capacity, overloading of the healthcare system, large displaced populations, as well as large populations suddenly without a source of income. These disasters require international assistance, through organizations such as the International Federation of the Red Cross and Red Crescent Societies (IFRC) and the United Nations (UN) cluster system.

IFRC is an umbrella organization that coordinates the activities of the 188 national societies in the Red Cross movement. In the event of a crisis, the national societies act to support the public sector organizations as distributors for relief, which is collected by the international mother organization from the global donor community (IFRC 2000).

The UN cluster system is organized around nine high-level types of humanitarian aid, referred to as clusters, each led by a UN agency and all coordinated by the Office for Coordination of Humanitarian Affairs (UN OCHA). The clusters with their lead agencies are Nutrition (UNICEF),

Health (WHO), Water/Sanitation (UNICEF), Emergency Shelter (UNHCR/IFRC), Camp Coordination/Management (UNHCR/IOM), Protection (UNHCR/OHCHR/UNICEF), Early Recovery (UNDP), Logistics (WFP), Emergency Telecommunications (OCHA/UNICEF/WFP) (United Nations 2013).

OCHA has a key role in the information management of the cluster system, as it is responsible for bringing together humanitarian actors to ensure a coherent response to emergencies. OCHA also ensures there is a framework within which each actor can contribute to the overall response effort. It is a top-level coordinating body for the humanitarian cluster system that deploys only during the most severe humanitarian crises. Both OCHA and the Red Cross conduct needs assessments before any intervention takes place. These products are then used throughout the cluster system to plan and inform response activities.

The remainder of this section is a summary of the information collection strategies of OCHA and the Red Cross, as outlined in (Inter-Agency Standing Committee 2012; IFRC 2000; Gralla, Goentzel, and Valle 2013). Situational assessment reports ("MIRA reports") created by OCHA are public and can be accessed online for most major humanitarian disasters in recent years.

2.4.3.1 Timeline of information needs

In the immediate aftermath of a rapid onset disaster, both OCHA and IFRC work with local teams to conduct needs assessments in the affected community, to know what kinds of aid to provide, where, and how or through whom to provide it. During the **first hours** after a disaster strikes, assessors will focus on identifying:

1. the presence, type, source and magnitude of the disaster;
2. an estimation of the number of casualties reported;
3. approximate number of properties damaged and the type of damage;
4. any immediate emergency priority needs, e.g. search and rescue, or first aid.

During the **first** days into the disaster, the focus will shift towards developing a plan to guide the intervention in the coming weeks or months. Development of such a plan requires identifying:

1. the full geographic and humanitarian extent of the crisis;
2. a detailed assessment of the affected community's needs, including the number of affected people, how they are affected and their demographic profile;
3. any implications for logistics caused by destruction or still-active threats;

4. ongoing response activities and relevant contact persons;
5. additional national and international response capacities;
6. gaps between needs (2) and interventions (4-5).

Gradually, comprehension of the disaster stabilizes and the needs assessment will further transition to track changes in needs, response activities, and available resources. Depending on the type of disaster, it may be relevant to look for new problems, e.g. aftershocks, looting or disease outbreaks. If necessary, more long-term intervention plans are established for the coming months.

In disasters where new hazards appear throughout the duration of the crisis, such as conflict, it is more difficult to define when in time particular assessment tasks needs to be supported. Change tracking – the third assessment stage – needs to include detection of new threats, and any resulting population movements, logistical implications, and humanitarian needs.

New methods proposed for data collection are valuable to international organizations if they can i) shorten the time required to carry out the first two stages of the needs assessment; ii) improve the quality of any of the assessments; or iii) reduce the cost of conducting any of the assessments.

2.4.3.2 Information structure

The situational assessments performed at the international level are used by decision makers with a wide range of responsibilities. Therefore, significant effort should be made to disaggregate the information along several dimensions, to enable filtering, aggregation and other types of dynamic reporting. The granularity of information required across each of these dimensions is dependent on the conditions of the specific disaster, but some guidelines are available and this section will attempt to give a structured overview of the information collected during a situational assessment.

During the course of the disaster, information will be collected not only about damage, but also about population movements, requests for resources or medical assistance, ongoing and planned intervention efforts, effects of interventions, and donation offers. These **report types** need to be kept separate, to allow comparison for instance of the magnitude of needs in relation to response efforts.

Both needs and interventions can typically be classified into one or several **sectors**. Typically, these are water, sanitation, health (physical and psychological), food, housing/shelter, displaced persons, clothes, money/income, security (legal), security (violence), education, infrastructure

(schools, health centres, houses/buildings, water points, roads/bridges, other), and access to information. Not all sectors are relevant in all disasters, and sub-categories may be introduced depending on the characteristics of the crisis. For instance, reports of violence may be categorized with higher granularity during a civil war than during the aftermath of an earthquake.

It is also important to track the activities, capacities and sometimes contact details of **involved entities**, in particular for response efforts. Entity relationship networks are also a key component of conflict analysis (McNaboe 2013). This enables the creation of so-called 3W databases, which record who does what where. Involved entities are also of key importance in man-made disasters, such as conflicts, where the actions and interactions of different actors are monitored for accountability purposes.

Typically it is not the individual reports that are of interest, but rather the patterns that emerge through aggregation of many reports. For instance, an analyst may want to compare disaster impact between geographic regions, or over time. This requires **quantifiers**, such as number of incidents, numbered of injured, number of casualties, number of displaced persons, number of active response initiatives, number of people helped, amount of resources donated, or amount of resources delivered. Furthermore, it is crucial that any duplicate or overlapping reports are detected and handled appropriately, to avoid that numbers are excessively skewed by counting the same incident multiple times.

For high-level aggregate reports used by coordinating staff, a **time** resolution of days or even weeks or months is commonly enough. Comparisons of key indicators before versus after the disaster struck are of particular interest, as are quantified time series of needs and interventions by sector. Some information may be shared with operational staff, such as search-and-rescue teams, or analysts wishing to piece together a timeline of events, in which case more detailed timestamps are needed, down to hours or even minutes.

Knowledge of the **geographic location** of events is critical, but it is difficult to define in advance at what geographic resolution each type of information needs to be presented. As disaster reports frequently use data aggregated all the way from national level, to provincial, district, neighbourhood and even building level, the only applicable advice for system developers is to store data at the highest geographic resolution available. In addition to the location itself, analysts may want to use other geographic datasets to segment the data, for instance to compare the effects of a disaster between urban and rural, mountainous and riverine, or high-income and low-income areas.

Finally, assessors will look for indications that particular **vulnerable groups** may be affected more strongly by a disaster than others. Depending on specific disaster conditions, it may be relevant to give special attention to reports relating to children, elderly, refugees, disabled, food-insecure, unemployed, ethnic minorities, or women.

While the information requirements differ between disasters, there are clearly similarities between disasters of similar types. It is therefore likely that predefined information templates could be developed for software systems targeted at supporting information management in disasters. In particular, sectors of interest are likely to differ between disasters of different types.

Needs assessments by international organizations also incorporate baseline data, and information from secondary sources, such as satellite imagery, country profile databases, and media reports. See page 13 of (Inter-Agency Standing Committee 2012) for a more extensive description of how external data is used.

2.4.3.3 Information output

The general goal of a needs assessment is to match the needs of an affected population against ongoing and available response capacity. This results in the identification of gaps, leading to strategic priorities. The needs assessment also supports the international organizations' other roles, such as advocacy, strategic decision making, interagency coordination, joint planning, fundraising resource allocation, monitoring, and evaluation.

While many ad-hoc techniques are used for reporting during disasters, there are some standards that can be followed. In particular, OCHA has developed the MIRA framework, consisting of a Preliminary Scenario Definition, a MIRA Report, and a Dashboard, each which could influence presentation techniques also in other information management systems. Finally, while mapping products and other reports are produced and published in digital format, they are often designed for print, to be easy to distribute, carry and share in the field.

2.4.4 Online volunteers and volunteer technical community

The past decade has seen the introduction and increasing impact of two new stakeholders in the disaster space – online volunteers and members of the volunteer technical community.

The functional role of both these groups in disaster response is to improve collective information management. Online volunteers have in numerous disasters proven to be a cheap, rapidly deployable and scalable human computation resource, consisting of hundreds of internationally distributed individuals, generally coordinated through membership of some volunteer organization. Notable examples are the Stand-By Task Force and UN Volunteers. Common tasks have included annotating, translating, organizing and summarizing social media content, photos and other media emerging from a disaster site; cleaning datasets; donor-recipient matchmaking; and raising international awareness of needs (Meier 2011; Harvard Humanitarian Initiative 2011; IFRC 2013c).

Like other disaster volunteers, to participate and to do so effectively, their online equivalents require awareness of the disaster; awareness of a volunteering opportunity; knowledge of how to carry out the work; personal motivation; coordination; feedback and psychological support; and tools to support the task at hand. Information requirements depend on the assigned task, from simple relevance classification of images that requires minimum external information, to more complex tasks such as trying to verify or refute a rumour, or to track down the location of a landmark referred to in a citizen report (Rogstadius et al. 2013b).

The tools that enable distributed online information management have so far been relatively experimental. Like all software development, the success or failure of these projects depends to a great extent on the developers' ability to define clear use cases, based on bottlenecks, breakdowns and unmet needs identified in the information flow within or between the different responder groups. Furthermore, developers need an in-depth understanding of user requirements and constraints imposed by disaster settings, such as data standards, common visual representations, available internet connectivity, bandwidth, communication devices, expected cognitive load, typical decision making tasks, and organizational workflows.

An overview of relevant systems and their features is provided in chapter 3, and a novel system, CrisisTracker, is presented in this thesis.

2.4.5 Media

As the name implies, mainstream media such as local, national and international radio, television and newspapers acts as an information mediator in disasters. Media collects, aggregates, prioritizes and redistributes information (Vasterman, Yzermans, and Dirkzwager 2005) to communicate information from and to all stakeholders, including affected populations, the media's own reporters, public sector organizations, international organizations, donors and NGOs. Reports in mainstream media are used by competent authorities to publicize warnings and guidelines (Vasterman, Yzermans, and Dirkzwager 2005), which raises local awareness of the disaster and causes potentially affected residents to seek out additional information and initiate volunteer response efforts (Rogstadius et al. 2013a).

Media also raises remote awareness about what the affected population faces, which can act as a trigger to initiate donor activities, aid operations to deploy or rebuild emergency infrastructure, and to mobilize non-local volunteers. However, it is not uncommon that media gives inaccurate or biased portrayals of disaster needs, or of incidents and actions during conflict or political unrest, which can cause improper allocation of aid resources (Jakobsen 2000). Furthermore, there is commonly a selection bias in the form of greater coverage of more sensational stories (McCarthy, McPhail, and Smith 1996). Due to the wide reach of mass media, these stakeholders are uniquely positioned to counteract the spreading of rumours. However, media

exaggerations can also increase fear and anxiety (Vasterman, Yzermans, and Dirkzwager 2005), and understatements can slow down response.

To be able to aggregate and publish information in a format that can be consumed by readers, listeners and viewers, media services first need a deep and broad awareness of ongoing events. The information needs are directly determined by the interests of the final information consumers, thus includes the humanitarian impact of the disaster (e.g. infrastructure damage, humanitarian needs of victims, the identity or numbers of displaced persons, casualties and injured); specific demographic groups in greater need; remaining or anticipated risks, threats and hazards; early reports of new events; ongoing and planned interventions; and in general changes to any long-lasting status. The collection methods employed include citizen interviews, reports submitted by citizens, spokesperson interviews, regular status updates released by public sector and international organizations, and salaried and freelance photographers.

All this information is processed, condensed and released in the form of news articles and reports characteristic to the particular media channel. Derived summaries can include top hazards, threats and risks; the greatest needs for intervention; the safety level of affected areas, which helps victims make decisions regarding evacuation from or returning to their home; damaged key infrastructure, e.g. transportation, utility lines and water quality, and estimations for when services will return; areas of donation needs; and ongoing and planned relief efforts. Local media also publishes warnings and guidelines issued by competent authorities.

2.4.6 Other stakeholders

In addition to the stakeholders mentioned so far, several other groups have key roles in the successful recovery after a humanitarian disaster. However, it has not been possible to find sufficient source material to reliably describe the information needed to support the decision making of these stakeholders. Future work in this area would be of great value to the volunteer technical community.

Domestic and international military forces play an increasingly important role, in countries across the entire scale of economic development. Military forces have several capabilities suitable for disaster response. These include special skills (transportation, urban search and rescue, mobile hospitals, surveillance and reconnaissance, situation assessment, damage assessment, and radiation monitoring), communications (equipment and trained personnel), and organized forces (equipment and disciplined personnel). Medical, transportation and communication capacities are particularly straight-forward to deploy in support of other stakeholders (Schrader 1993; Wiharta et al. 2008; Hofmann and Hudson 2009; Brattberg 2013). In man-made disasters, e.g. conflict, military and other armed forces play a far more complex role. From a disaster

management perspective, military as well as rebel and mercenary forces may for instance constitute a direct threat to affected populations, they may provide armed escort for aid operations, and provide surveillance capabilities.

Private sector for-profit companies also play an increasingly important role in disaster response and recovery. Examples include the provision of post-disaster satellite imagery, heavy machinery, logistic services, medical assistance, IT infrastructure and communications, and management skills. Private sector non-profit organizations, or non-governmental organizations (NGOs), provide several other important functions in the recovery process. Common roles include donor management, distribution of donations, provision of shelters, management of refugee camps, medical assistance, volunteer management, and mapping services.

Finally, international, domestic and local donors, including individuals, funds, organizations and governments, contribute vast financial and material resources needed for the recovery. At a basic level they require knowledge of local needs and distribution channels, both to be able to donate, but also to avoid donating unnecessary or redundant resources that clog up the distribution chain (Pan American Health Organization 2000).

2.5 Summary

Response to and recovery from humanitarian disasters is handled through organized efforts by local, national and international emergency response organizations, combined with the collective efforts of regular citizens acting as independent spontaneous volunteers. Both the effectiveness and efficiency of the response can be improved through increased situational awareness. Basic perception of the situation can be gained through information collection, and comprehension, projection and prediction capacity comes with increased experience, training and system support.

Information that contributes to situational awareness is in practice defined by the context and goals of a specific situation. While several stakeholders have information needs specific to that group, several information types are useful across the responder spectrum. However, while the information itself can be shared, the resolution at which information needs to be presented differs with the scope of the decision maker's responsibilities. An analyst in an international organization, a local organization distributing food and water, and a resident of a disaster affected community all share a common interest in the locations and nature of hazards, needs and intervention efforts, but they are unlikely to be helped by the same maps and reports. This suggests a potential great value of information management solutions that share a common high-resolution model of the disaster, on top of which several views are provided, each targeted at a specific decision making role.

Though social media is a relatively novel phenomenon, it can be seen in its historical context merely as a communication channel through which disaster affected people, at any level, exchange information relevant to decision making. Content analysis of Twitter status messages confirms that large volumes of messages are posted on social media during disasters, which contain information that is relevant to decision makers in all stakeholder groups that this chapter has reviewed. Unique to social media, in particular Twitter, is that much of the communication is publicly accessible through APIs, which, at least in theory, makes it possible to monitor the communication centrally to leverage the online community as a distributed sensor network.

Despite this, significant challenges remain before this content can be utilized in decision making. Social media has a low signal-to-noise ratio, with non-informative messages greatly outnumbering informative content. Furthermore, during the greatest disasters, even the informative messages alone can be expected to be far too numerous to all be absorbed by a user of any system. The next chapter gives an overview of state of the art approaches to handling these issues, and outlines the direction for research presented later in the thesis.

Chapter 3

Information management approaches

Collecting and displaying situational awareness-related information in real-time is imperative during humanitarian disasters. As the previous chapter showed, existing content available on online social media is an attractive new source of information, as it can potentially allow interested stakeholders to leverage the vast user base of these platforms as a distributed sensor network. This can, in theory, allow rapid, reliable and cheap detection of new events, with access to rich evidence such as videos and images, and even enabling direct communication with eyewitnesses. Furthermore, even if for technical reasons the tools may not be usable during the immediate onset of a disaster, they can serve an important sociological function once infrastructural repairs are made (Palen and Liu 2007).

Several processing techniques and software systems have been developed to aid in this, or similar, information management processes. This chapter presents a comparison of theory and systems that can be loosely grouped into machine-based information extraction (Kumar et al. 2011; Yin et al. 2012; Abel et al. 2012; Best et al. 2005) and crowdsourced human-based computation (Ushahidi). The remainder of this chapter will examine the specific features of these systems, summarized in table 3.1, and discuss the strengths and weaknesses of machine- and human-based information processing.

The review identifies three key qualities of the processing techniques – scalability, accuracy and flexibility – which need to be combined for a system to realistically improve situational awareness of disaster responders based on content mined from online social media. Currently, none of the available systems score highly in all three of these qualities. The chapter thus identifies areas where further research is needed and proposes how future solutions can combine ideas from several applications into more effective, efficient or complete solutions.

	News-Brief[2]	Twit-cident[3]	Tweet-Tracker[4]	Yin et al.[5]	Twitris 2.0[6]	Usha-hidi[7]	Crisis-Tracker[8]	CT + AIDR[9]
Machine processing	●	●	●	●	●	○	◐	●
Human-based computation	○	○	○	○	○	●	●	●
Designed for social media monitoring	○	●	●	●	●	◐	●	●
Designed for tracking disasters	○	○	●	◐	○	●	●	●
Easily adaptable, e.g. to new event types and languages	○	◐	○	◐	○	●	●	●
De-duplication of redundant reports	●	○	○	●	◐	○	●	●
Summarization of related reports	●	○	○	◐	○	○	●	●
Content ranking to prevent information overload	●	◐	◐	●	◐	○	●	●
4W extraction (who, what, where, when)	●	●	○	◐	◐	●	●	◐
4W filtering in real-time	●	●	○	◐	◐	○	○	◐

TABLE 3.1: Feature comparison of related and proposed systems.
● Supported; ◐ partly supported; ○ not supported.

3.1 Machine-based information extraction

Machine-based information extraction systems build on decades of research in data mining, machine learning and natural language processing (NLP) to perform information extraction (e.g. named entity recognition, sentiment analysis, and entropy- or network-based ranking), data reduction (clustering or filtering), summarization and visualization. The automation makes it possible to process very large data volumes in short time, which is necessary for real-time streaming social media. Machine-based aggregation of mainstream news media has been a great success, with consumer applications like Google News[1] now available on the web and most consumer smart phones.

[1] https://news.google.com/
[2] (Best et al. 2005)
[3] (Abel et al. 2012)
[4] (Kumar et al. 2011)
[5] (Yin et al. 2012)
[6] (Jadhav et al. 2010)
[7] http://www.ushahidi.com/
[8] Chapter 6
[9] Chapter 7

This research has so far almost exclusively focused on document collections such as news articles, books, scholarly articles and web pages, and existing algorithms can roughly be classified into one of two families. The first treats documents as unordered generic "bags of words" which are grouped, ranked or classified based on the frequency with which words occur locally within documents and globally between documents. The second family uses a range of natural language rules and patterns to extract terms and semantic relationships that represent the underlying "meaning" of the text.

In contrast, citizen-generated information collected from social media is typically in the form of very short texts (less than 140 characters for some media), images and video, which is too sparse for most statistical techniques to work well, or is not of textual nature at all. The textual content is often also authored using relaxed rules for spelling and grammar (e.g. *"New threat to #Syrian #refugees #polio http://t.co/9pGD4PIqLp #cnn #Syria"*) and, of particular relevance in disaster settings, is often in a local language other than English, for which the majority of NLP-research has been conducted. In addition, document collections from social media arrive as streaming data and the number of documents can be several orders of magnitude greater than those seen in traditional document collections. This essentially restricts the solution space to online language-independent algorithms operating in linear time with regards to the number of documents. Methods of higher complexity order can be used if applied iteratively over different segments of the document stream, but this approach comes with undesired processing delays and the output of different runs of an algorithm cannot always be merged.

As discussed in the previous chapter, meta-data such as needs sector, geographic location, named entities and report type play an important role in being able to contextualize, filter and aggregate information to match the information needs of specific decision makers. If such meta-data cannot be reliably extracted, it consequently becomes very challenging to design user interfaces which present information relevant to a specific user role. Despite these challenges, a number of noteworthy systems have been built using machine-based processing techniques.

EMM NewsBrief[10] (Best et al. 2005) automatically mines and clusters mainstream news media from predetermined online sources in a wide range of languages, with new summaries produced every ten minutes. It relies on rule-based classifiers for meta-data extraction, and substantial investment has been made over more than a decade to create such rules for a wide range of languages. Despite this great investment, the system has not been extended to handle social media. EMM NewsBrief ranks news stories by their global coverage; the more a story is reported, the more important it is deemed to be. The approach works well for detecting globally impactful events, but comes with a time-lag as a story needs to gain sufficient traction before it is highlighted by the system. Users of the system can also filter stories based on the different dimensions of extracted meta-data, to better match their interests. While the system cannot

[10] http://emm.newsbrief.eu/

be used for social media monitoring, many of the ideas behind EMM NewsBrief have influenced the work presented in this thesis.

Several systems (Kumar et al. 2011; Yin et al. 2012; Abel et al. 2012; Jadhav et al. 2010) have been proposed that parse a Twitter feed to extract and rank popular terms, user mentions and URLs, to display maps for geotagged tweets and users, and word clouds for popular terms. The systems then focus on providing quantitative aggregate metrics of message counts, such as line graphs displaying the number of tweets mentioning a particular term over time, maps showing the source locations of collected tweets, or retweet networks visualizing how information has propagated. Users can then select time periods, terms, or locations for which to display a list of matching messages. Twitris 2.0 (Jadhav et al. 2010) integrates a general natural language processor developed for news articles. The same is true for the Twitcident (Abel et al. 2012) system, which further allows users to filter content based on the categories in its ontology, as well as adding new categories through manually defined classification rules. Yin et al. (2012) use a series of hierarchical word clouds to support multi-level content drill-down, and provide pre-trained English-only classifiers to detect messages mentioning infrastructure damage and a few other categories. As evident by table 3.1, this system shares many features with the solutions proposed later in this thesis. The primary way in which the two solutions differ is that Yin et al. (2012) focused more strongly on the detection of new emergencies, while this research has placed greater emphasis on maintaining situational awareness during long-lasting complex disasters.

Common to all these systems is that no evaluation has taken place of their performance in improving situational awareness of users during ongoing disasters. Furthermore, none of the systems have to the author's knowledge gained much traction among end users, making it difficult to reliably assess their practical utility in supporting situational awareness and decision making during disasters. However, based on the information needs identified in the previous chapter, the value of the provided features is likely to be fairly limited. For instance, maps displaying the source locations of messages are likely to highlight locations where infrastructure is intact and where population density is high, while needs assessment maps should highlight those areas that are most severely affected by a disaster. Word clouds displaying single terms strip out the context in which those terms occurred, and it is unclear if de-contextualized sets of automatically extracted terms can at all provide the form of meaningful navigation and filtering that would be provided by sector-based classes such as water, health, or shelter. Information propagation networks may have applications for assessing the credibility of a specific claim, but are unlikely to improve overall situational awareness. Finally, there are indications that classifiers trained on social media content collected during one disaster may generalize poorly when applied to other disasters (Imran et al. 2013b).

In part, these are likely to be design issues caused by insufficiently documented use cases and information needs, but the systems also illustrate limitations in what current computational techniques alone can achieve.

3.2 Crowdsourced human-based computation

Systems for crowd reporting are built around human-based computation, which is a computer science technique in which a computational process performs its function by outsourcing certain steps to humans. This is often done in a crowdsourced manner, where the computation is done by a distributed group of people. The technique is practical for handling computational problems that are easy to solve for humans, but challenging for computers, such as image labelling (Ahn and Dabbish 2004), audio transcription, knowledge management (Kuznetsov 2006) and solving business problems (Vukovic 2009). The performance of crowds largely depends on incentives (Rogstadius et al. 2011b) such as money (Kaufmann, Schulze, and Veit 2011), game mechanics (Ahn and Dabbish 2004; Cooper et al. 2010), social capital (Raban 2008) and public good (Kuznetsov 2006).

Crowdsourcing in the disaster domain has primarily been carried out through the Standby Task Force (SBTF 2012), a volunteer-based network of crisis mappers established in 2010. Since its creation, the special-purpose group has grown significantly and now represents a source of motivated crowds willing to deploy for a variety of information collection and processing tasks during humanitarian disasters. Specialized online market places for crowdsourcing also exist, such as Amazon's Mechanical Turk, where workers can carry out tasks lasting a few minutes or seconds in return for a small monetary reward. For-pay crowdsourcing markets have not been used for real-time human computation in disasters, and could be an alternative and highly scalable approach to handle spikes in information volumes, before volunteers can be recruited through other channels. Furthermore, it is possible that the great intrinsic value of work during disasters could compensate for the need to pay workers even on these markets, but it is unclear how this would affect recruitment and accuracy of workers.

Human-curated crisis maps are a central form of online volunteering that began emerging after hurricane Katrina in 2005 (Palen and Liu 2007). Crowd-reporting systems like Ushahidi[11] and Google Maps[12] combined with Google Docs[13] enable curation and geo-visualization of manually submitted reports from a wide range of sources.

In Ushahidi, user-submitted reports can be annotated with report category and geographic location, which then enables visualization of the geographic distribution of reports in each category.

[11] http://www.ushahidi.com/
[12] https://maps.google.com/
[13] https://docs.google.com/

The platform was first deployed in 2008 to allow Kenyans to report human rights violations during post-election unrest (Harvard Humanitarian Initiative 2012). Witnesses submitted reports via web-forms, email and SMS and mainstream media reports were also mapped. Notable Ushahidi deployments have since been launched in Haiti, Chile, Pakistan, Russia, Syria, Tunisia, Egypt, New Zealand, Sudan, Libya and Somalia, many on request by the UN or other international organizations, though most are set up by ordinary individuals (Meier 2011). The platform is also often used as a repository for information manually collected from social media and other channels.

Due to reliance on human processing in all information-processing stages, Ushahidi's effectiveness depends entirely on the size, coordination and motivation of crowds. The majority of the most successful deployments have been powered by the volunteer-based Standby Task Force (SBTF 2012), which has set up dedicated teams for media monitoring, translation, verification, and geo-location. This team structure is further supported by task management extensions[14] to the platform and adapts well to needs of specific disasters, but current tools still offer insufficient support even for dedicated crowds to cope with the torrent of information posted on social media during mass disasters (Meier 2012). Large volunteer groups are also difficult to sustain over longer periods of time, with the longest SBTF deployments lasting only four weeks (Meier 2011).

Evaluation (Morrow et al. 2011) of a landmark deployment of the Ushahidi deployment following the Haiti earthquake in 2010 showed that the platform filled several information gaps, in particular during the first days and weeks, before the UN and large organizations were fully operational. Strategic, operational and tactical organizations used the map, integrated with other traditional information sources, to develop an assessment of the situation on the ground. The open platform also provided situational information for small NGOs that lacked field presence, and for affected locals, the Haitian Diaspora and private relief coordinators. According to the evaluation, the Ushahidi platform provided situational awareness and critical early information with geographic precision that is lacking in other situational awareness tools available to the public.

However, the evaluation also found that decision makers questioned the way information was classified and organized in the system, pointing out a disconnect between the classification scheme and organizational information needs and established data structures. Moreover, many questioned how representative the information was of the situation on the ground, given that only a few percent of collected reports were human-processed and thereby made available to end users.

[14] http://roguegenius.com/

	Scalability	Accuracy	Flexibility
Machine-based information extraction	Full automation allows real-time stream processing of content at the rate it is generated by online social media users.	Simple keyword filtering and classification available. Not capable of assessing relevance, importance and implications of news, images and video, in relation to current humanitarian goals.	Incapable of solving any problem not considered at system design time. Generally struggles with new natural languages, new media types and situation-based information needs.
Crowdsourced human-based computation	Cannot keep up with live content streams from online social media during mass disasters. Management capacity has proven difficult to scale along with crowd size and disaster magnitude.	Human workers are naturally adept and comprehending information published on social media. Crowdsourcing techniques can be used to verify and aggregate the output of individual workers.	Highly flexible. New media types and analysis tasks can easily be learnt by human workers. Processing of new content languages can be handled through recruitment of new workers.
Hybrid processing	Repetitive steps are fully automated, e.g. data collection and some metadata extraction. Work items are assigned to human workers by a priority ranking metric. Machine-learning techniques are integrated to enable the system to apply human processing behaviour at scale.	Machine processing is used whenever it is of sufficient quality. Human processing is applied to processing tasks which cannot be completed computationally. Which tasks these are may vary by situation.	The system can learn processing patterns from examples provided by human workers to adapt to new situation-based processing requirements. Human processing is always available as a fall-back.

TABLE 3.2: Strengths and weaknesses of the three information management approaches.

3.3 A hybrid approach

The ultimate purpose of information management in crisis is to provide better situational awareness so as to make more informed decisions regarding possible interventions. To fulfil this purpose, information management systems must be capable both of separating signal from noise, and to identify particularly valuable pieces of information in large filtered sets of reports that all relate to decisions facing target user individuals or organizations. Information should be provided that is not only relevant topic-wise, but also at an appropriate spatial and temporal resolution. System components for information processing, visualization and analysis should all be designed to help users gain situational awareness at perception, comprehension, projection and prediction levels.

However, with the current state of the art in information processing methodology, it is unlikely that these requirements can be met using only automation or human-based computation alone. Algorithmic approaches can be scaled to handle the torrents of information generated during disasters, but are not accurate enough to extract information needed for decision making, nor sufficiently flexible to meet unanticipated processing needs in new disasters. Conversely, crowdsourced human-based computation is accurate and flexible, but very difficult to scale. These and other strengths and weaknesses discussed in this chapter are summarized in table 3.2.

This thesis will investigate if a hybrid approach (Rogstadius et al. 2011c), in which human-based and machine computation work hand in hand to complement each others' weaknesses, can cope with the many challenges imposed by the application domain and data source. The vision is set at processing and summarization on par with EMM NewsBrief, but for citizen-generated content published on social media during an ongoing disaster. This requires finding adequate approaches for breaking news detection, aggregation or summarization of content in many messages, ranking of aggregate stories, extensive meta-data extraction for navigation, together with management strategies for maximizing the performance of crowds.

Through a series of studies, different components of a hybrid system are developed, which are together integrated in the CrisisTracker system. Chapter 4 investigates if the intrinsic value of work carried out for humanitarian purposes can compensate for no or lower payment if work is submitted to for-profit crowdsourcing markets. Chapter 5 proposes to use online text clustering to handle many of the challenges in processing social media streams, and presents an initial feasibility study based on exploratory analysis of five clustered social media datasets. Chapter 6 presents the architecture of a basic hybrid system, along with evaluation of the system in a conflict monitoring scenario in terms of timeliness and utility compared to established information collection methods. Chapter 7 further explores hybrid information curation, presenting the architecture and initial evaluation of a classification module that uses supervised learning to generalize the tagging behaviour of human curators. Finally, the work is tied together and conclusions and directions for future work are presented in chapter 8.

Chapter 4

Feasibility study of humanitarian crowdsourcing on for-pay task markets

4.1 Introduction

Crowdsourced human computation is a powerful approach to handling problems that by nature are difficult to solve computationally. The method is analogous to parallelizing computational work in programming environments and typically consists of segmenting the work into multiple small and independent pieces, which are then dispatched along with instructions through a crowdsourcing system to be solved by humans. In the context of humanitarian disasters, crowdsourcing can be utilized by an organization or as part of a system to in a centralized manner issue micro-tasks that need to be completed. Such tasks can be, for example, analysis of satellite imagery, disambiguation of incoming information, and collection of relevant information.

Especially interesting forms of crowdsourcing are general-purpose task markets such as Amazon's Mechanical Turk (MTurk), in which a variety of different tasks can be posted. Popular crowdsourcing tasks include image tagging and classification, audio transcription and various types of surveys, all of which could find applications in disaster settings. However, while recent community-driven ICT responses to emergency events have mostly relied on individuals' willingness to contribute time and effort for free, this is not the case on crowdsourcing markets where workers expect to be rewarded for their time, and in fact develop strategies for optimising their rewards. An important question, therefore, is whether individuals on for-pay crowdsourcing markets can be motivated to donate their time for an apparently worthwhile cause such as a non-profit charity.

Furthermore, crowdsourcing work involves a number of challenges different from those faced in traditional work settings. Crowd workers in general purpose markets like MTurk may have

highly varying expertise, skills, and motivations. Employers ("requesters" in MTurk) have very little visibility into these characteristics, especially compared to a traditional organization in which workers are vetted during recruitment, have work histories, have reputations within and outside the organization, and may go through organizational socialization methods such as training to ensure they can appropriately satisfy their job requirements. Furthermore, workers can easily return work for a given job with no repercussions or even create an entirely new profile with a clear reputation. These challenges mean that employers have to rely largely on motivational factors as a means of eliciting high quality output.

The study presented in this chapter, first published as (Rogstadius et al. 2011b), aims to shed light on this issue: are workers on a crowdsourcing website willing to donate their time for charity? In addition, is the quality of their work affected when doing work for charity, and if so how? Finally, are motivational factors such as meaningfulness and payment independent of each other, or are interaction effects present that need to be considered when crowdsourcing work in humanitarian settings? Answering these questions will contribute to an understanding of if and when crowdsourcing markets can be used for human computation during disaster response.

The study uses a novel experimental methodology that controls for self-selection effects, and a novel experimental task that allows for a wide range of participant accuracy.

4.2 Related work

A traditional "rational" economic approach to eliciting higher quality work is to increase extrinsic motivation, i.e., how much an employer pays for the completion of a task (Gibbons 1997). Some evidence from traditional labour markets supports this view: Lazear (2000) found workers to be more productive when they switched from being paid by time to being paid by piece; Hubbard and Palia (1995) found correlations between executive pay and firm performance when markets were allowed to self-regulate.

However, there is also evidence that in certain situations financial incentives may not help, or may even hurt. Such extrinsic motivations may clash with intrinsic motivations such as a workers' desire to perform the task for its own sake. For example, a classic experiment by Deci (1975) found a "crowding out" effect of external motivation such that students paid to play with a puzzle later played with it less and reported less interest than those who were not paid to do so. In the workplace, performance-based rewards can be "alienating" and "dehumanizing" (Etzioni 1971). If the reward is not substantial, then performance is likely to be worse than when no reward is offered at all; insufficient monetary rewards can act as a small extrinsic motivation that tends to override the possibly larger effect of the task's likely intrinsic motivation (Gneezy and Rustichini 2000b). Given that crowdsourcing markets such as Mechanical Turk tend to pay

very little money and involve relatively low wages (Ipeirotis 2010), external motivations such as increased pay may have less effect than requesters may desire. Indeed, research examining the link between financial incentives and performance in Mechanical Turk has generally found a lack of increased quality in worker output (Mason and Watts 2009)[1]. Although paying more can get work done faster, it has not been shown to get work done better.

Another approach to getting work done better could be increasing the intrinsic motivation of the task. Under this view, if workers find the task more engaging, interesting, or worth doing in its own right, they may produce higher quality results. Unfortunately, evidence so far has not supported this hypothesis. For example, while crowdsourcing tasks which are framed in a meaningful context motivate individuals to do more, they are no more accurate (Chandler and Kapelner 2013). In summary, no approach has yet found extrinsic or intrinsic motivations to increase the quality of crowd workers' output[2].

However, there are a number of issues that suggest the question of motivating crowd workers has not yet been definitively settled. First, prior studies have methodological problems with self-selection, since workers may see equivalent tasks with different base payment or bonuses being posted either in parallel or serially. Second, to our knowledge no study has yet looked at the interaction between intrinsic and extrinsic motivations; Mason and Watts (2009) vary financial reward (extrinsic), while Chandler and Kapelner (2013) vary meaningfulness of context (intrinsic) in a fixed diminishing financial reward structure. Finally, the task used in (Chandler and Kapelner 2013) resulted in very high performance levels, suggesting a possible ceiling effect on the influence of intrinsic motivation.

4.3 Crowdsourcing and Mechanical Turk

Amazon's Mechanical Turk (MTurk) is a general marketplace for crowdsourcing where requesters can create Human Intelligence Tasks (HITs) to be completed by workers. Typical tasks include labelling objects in an image, transcribing audio, or judging the relevance of a search result, with each task normally pay a few cents (USD.

Work such as image labelling can be set up in the form of HIT groups, where the task remains identical but the input data on which the work is carried out varies. MTurk provides a logical workflow within such groups where workers are continuously offered new HITs of the same type

[1] The relationship between price and quality has also had conflicting results in other crowdsourcing applications such as answer markets (Harper et al. 2008).

[2] Though there are other methods; for example, (Kittur, Chi, and Suh 2008) used a variety of methods to increase signal in subjective tasks, such as signalling monitoring or increasing the cost of bad faith answers. Another example is CrowdFlower's "gold standard" approach, which provides feedback to workers when they answer specific sampled questions incorrectly. However, these are task-specific approaches that may not work for many kinds of tasks, and while they may filter out poor quality work by raising the threshold for acceptance, may not motivate high quality output.

after they accept and complete a HIT within the group. MTurk also allows splitting a HIT into multiple identical assignments, each which must be taken by a different worker, to facilitate for instance voting or averaging schemes where multiple workers carry out the same task and the answers are aggregated.

4.4 Running controlled studies on Mechanical Turk

Using MTurk poses a problem for experimental studies, since it lacks support for random participant assignment, leading to issues even with between subjects control. This is especially problematic for studies of motivation, as self-selection is an inherent aspect of a task market. This means that results in different conditions could be due to attracting different kinds of people rather than differences in the conditions themselves. In this study, given two tasks of which one pays more and one pays less, making both of them available on the site at the same time would bias the results (contrast effect)[3]. If they were put up at different times, then different workers might be attracted (e.g., Indian workers work at different times than Americans; some days/times get more activity than others, etc.), or more attractive work could be posted by another requester during one of the conditions but not the other.

The other extreme is to host everything on the experiment server, using MTurk only as a recruitment and fulfilment host. All participants see and accept the same identical task, and are then routed to the different places according to the appropriate condition on the experimenter's side. This fails when studying how workers act naturalistically, as everything is on the host environment. Thus aspects such as the title, description, and most importantly reward cannot be varied by condition, making it impossible to study natural task selection.

This study proposes a novel approach in which participants fill out a common qualification task with neutral title and description. This qualification task (in our case, simply collecting demographic data) is hosted on the experimenter's server and on completion randomly assigns the participant to one of the conditions through a condition-specific qualification in the MTurk system. This qualification enables workers to see and select only tasks in that condition when searching for tasks in the natural Mturk interface. In this study we used an MTurk qualification type with six different possible values corresponding to the different conditions. The key benefit of this approach is that participants still use the MTurk interface as they naturally do to self-select tasks, which can have condition-specific titles, descriptions, content, and rewards. While participants can still explicitly search for the tasks in other conditions and see them in some HIT listings, HITs cannot be previewed without having the appropriate qualification. Hosting the task externally (which we did not do) would avoid the explicit search problem, but would

[3]This contrast effect would be problematic even for non-simultaneous posting if workers saw one task at one price and then the same task at another price at a later time.

not address non-preview textual descriptions or the key issue of supporting condition-specific variations in payment.

Another advantage of the qualification-task-approach is that the worker will always retain the qualification granted to them by the experimenter (so they can be kept track of). Thus, for example if an experimenter wanted to make a new experiment available to a subset of their participants they could add the qualification for it to the appropriate participants and the task would automatically become available to the target participants on MTurk. For more intensive recruitment, once a worker has completed the qualification task and their worker ID is known, they can be emailed directly by the experimenter, even if they did not complete an experiment. This proposed approach for recruiting participants from a crowdsourcing market lets us retain some of the control of a traditional laboratory setting, the validity of participants searching for work in their natural setting, and the benefits offered by a greater diversity of workers more representative of the online population than undergraduates would be (Horton, Rand, and Zeckhauser 2011). The legitimacy of doing both cognitive and social experiments with Mechanical Turk has been supported by multiple studies, e.g. (Heer and Bostock 2010; Ipeirotis 2010).

4.5 Study

With the goal of measuring the interaction effects of intrinsic and extrinsic motivation on Amazon's Mechanical Turk, we decided on a 2x3 design for our experiment. We implemented our motivation manipulation through two levels of a "cover story" (non-profit, for-profit, each described in more detail below) and three levels of reward (0, 3, and 10 cents USD). We then designed a task that allows us to quantitatively measure the quality of the work in a way where quality is dependent on effort while avoiding ceiling effects. Based on the results from previous work, we worked primarily with four experimental hypotheses:

H1 Tasks in the non-profit (i.e. charity) conditions will be completed faster than tasks in the for-profit conditions.

H2 Tasks in the non-profit (i.e. charity) conditions will be completed more accurately than tasks in the for-profit conditions.

H3 Tasks in high-pay conditions will be completed faster than tasks in low-pay conditions.

H4 Tasks in high-pay conditions will be completed more accurately than tasks in low-pay conditions.

4.5.1 Recruitment

To recruit participants, a Human Intelligence Task (HIT) was posted on Amazon's Mechanical Turk (MTurk), appearing to be from a fictitious organization that handles crowdsourcing on behalf of third party pharmaceutical and health-related organizations. The HIT advertised that by completing the associated questionnaire workers would obtain a qualification to complete further HITs. The HIT consisted of an externally hosted questionnaire that collected broad demographic data from participants, as well as data on their experience on MTurk. Once completed, the questionnaire allocated participants to one of the six experimental conditions by assigning them one of six different qualifications on MTurk, and in addition awarded participants a one-off bonus of 2 cents USD.

Upon completing the questionnaire and obtaining a qualification, participants gained access to further HITs in their assigned condition. These HITs could be accessed either through a link provided at the final confirmation page of the qualification form, through an email sent to them or through regular search.

A worker who would list all work currently available from the fictitious organization could at any given time see six HIT groups with generic and identical titles ("Medical image analysis") and descriptions ("See HIT preview for instructions") but with different payment levels and requiring different qualifications. However, workers listing work available to them would only see the HIT group relevant to their qualification (if any), and in any case could preview only the qualification-relevant HIT group to see a detailed description and image of the actual task.

On average, a little over a day elapsed between when working participants submitted the questionnaire and when they accepted the first task. However, there was a significant dropout effect in which most workers who went through the registration process (81.3%) did not complete any experimental tasks at all.

4.5.2 Experimental task

The experimental task consisted of a single HTML page that included a cover story at the top of the page, instructions on how to complete the task, an image to analyze, and input fields for answers. The cover story for the non-profit condition was:

*The **Global Health Council**, a non-profit organization and the world's largest membership alliance dedicated to saving lives by improving health throughout the world, is running a study to assess the effectiveness of recent advances in the treatment of malaria.*

Blood cell
COUNT these, including partially visible cells.
Malaria parasite in ring-form with double chromatin dots.
COUNT these.
Malaria parasite in other growth stage.
IGNORE these.

FIGURE 4.1: Instructions given to participants on how to complete the experimental task.

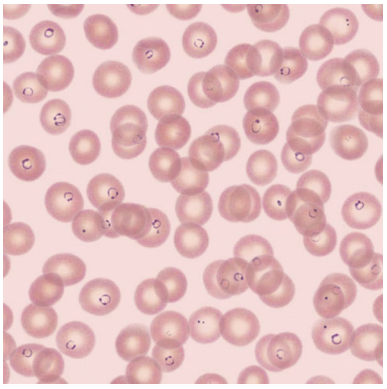

FIGURE 4.2: A sample image of medium complexity from the experimental task.

The for-profit statement gave the same information, except that the organization was changed to "***Rimek International***, *a major actor in private pharmaceutical manufacturing*". The instructions for all conditions then had as follows:

This task requires you to identify blood cells infected with malaria parasites. The malaria parasite goes through a number of growth stages. For this task you are required to identify the parasites that are in a specific growth stage (ring-form with two adjacent dots). Look at the image below and

1. *Count the number of malaria parasites in ring-form, having double chromatin dots.*
2. *Count the total number of blood cells in the image.*

Some images may be ambiguous and require guesses or estimates. Please keep in mind that the quality of any such estimates will directly influence the quality of this research.

The instructions concluded with a legend of objects featured in the task image (Figure 4.1).

After the instructions, participants were shown a computer-generated image with known properties (Figure 4.2), and were asked to enter i) the number of malaria parasites in the correct growth stage in the image, and ii) the total number of blood cells.

The experimental images were generated by independently varying the number of cells in the image, and the number of the malaria parasites that participants had to count. Around 18% of cells contained noise in the form of parasites in non-interesting growth states and images

with high cell counts had significant visual overlap of the cells. Initial tests showed that the experimental tasks would take between 30 and 200 seconds to complete, with an average of around one minute.

Upon completion of the HIT, participants were automatically given the option to complete further HITs in the condition. By clicking on the "accept" button, participants could attempt another HIT. Each condition consisted of 100 HITs, with two identical assignments per HIT.

4.6 Results

The study ran for 48 days. Once all the advertised HITs in a condition were completed no more participants were allocated to that condition. However, neither 0 cent conditions attracted enough workers for all tasks to be completed; an issue we will return to later. To minimize bias all conditions appeared to be available in the public listing, even though all work was completed for some conditions and they did not accept new participants. In a few instances (4.7%) assignment answers had been swapped for parasite and cell counts. These answers were manually corrected when the answers differed by more than 25% and when the error was lower after swapping.

The 3 and 10 cent rewards were based on an estimated average task completion time of one minute, which would have yielded hourly wages of $1.8 and $6 USD respectively. In practice, however, participants spent more time than estimated per task and achieved effective hourly wages of only $1.4 and $3.3 for the 3 and 10 cent groups.

4.6.1 Demographics

A total of 843 people completed the qualification questionnaire, of which 158 showed up, i.e. completed at least one assignment. Unless otherwise stated, these are the participants to which the results refer. Of the participants that showed up, 49% were female. In addition, 42% reported having lived only in South Asia (including China and India) and 35% only in North America (excluding Mexico). Participants from South Asia compared to those from North America on average had lower yearly income (median ¡$5k vs. $20k-$60k), higher education and were younger. The median working participant had a bachelor's degree and was 25-34 years old.

4.6.2 Metrics

For each experimental task (assignment) the following information was collected: reported cell count, reported parasite count, time spent and participant ID. In addition, for each participant the following information was recorded by the questionnaire: demographics (gender, age, education, income, region(s) of residence), time registered on MTurk, weekly time spent on MTurk, diseases affecting user or somebody close to them (including malaria), previous experience with blood analysis.

To measure the effort that each participant chose to spend, we use total completed assignments, total working hours and mean time per task. Uptake ratios (ratio of registering participants who completed at least one task) are also reported, as they have implications for total work completion rates.

An aggregate accuracy metric was defined to capture quality of answers as

$$accuracy = 1 - \frac{1}{2}\left(\frac{|p_{est} - p_{real}|}{p_{real}} + \frac{|c_{est} - c_{real}|}{c_{real}}\right), \tag{4.1}$$

where p is parasites and c is cells.

A combined metric for task complexity was also introduced, with greater weight given to parasites than to cells as participants had to consider the growth stage of the parasite when counting them:

$$complexity = c_{real} + 3p_{real}. \tag{4.2}$$

4.6.3 Work effort

Figure 4.3 shows the rate at which the assignments in each condition were completed, with higher rates for higher paying conditions. The progress rate is not steady since most progress comes in short bursts from single individuals who choose to complete many assignments in one go.

Most participants chose to complete only a few tasks, and the distribution was heavily skewed (mean 6.5, median 2). Figure 4.4 shows how the total workload was distributed among participants. The graph shows that a single participant contributed half the total work in the for-profit 0-cent condition, and that as much work was produced by a single participant in the for-profit 10-cent condition as was produced in total in the non-profit 0-cent condition. This distribution can be expected when workers are allowed to self-select how many tasks to complete and it is representative of normal work distribution on MTurk.

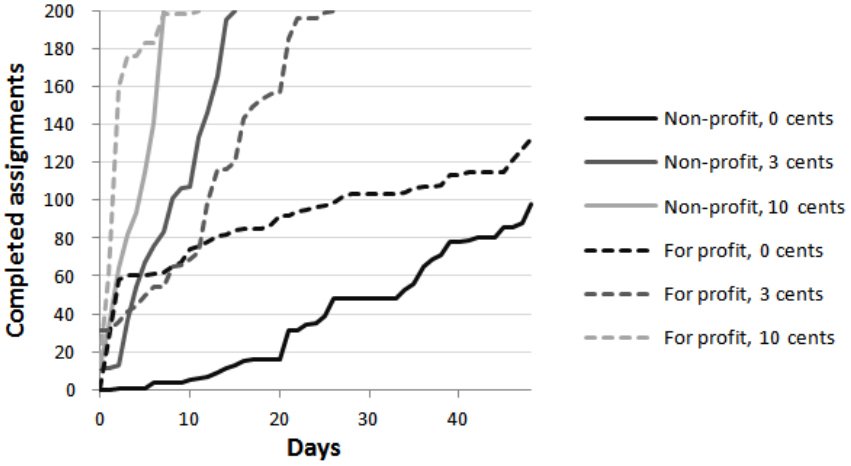

FIGURE 4.3: Time taken to complete each condition's batch of assignments. Contributions from individuals who chose to complete many assignments in a sequence show up in the graph as vertical jumps in the time series.

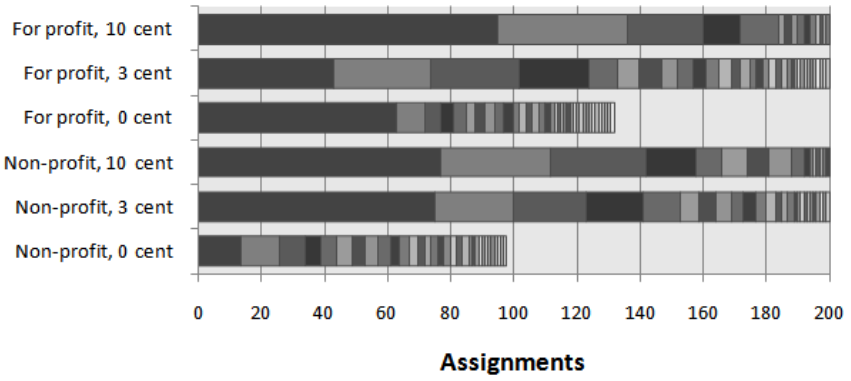

FIGURE 4.4: Distribution of completed assignments among participants. Each participant is represented by one bar segment. The two 0 cent workloads did not complete.

Table 4.1 lists various indicators of interest, including uptake (percentage of registering participants who completed at least one assignment). Payment variations had clear effects, with total uptake numbers of 12.9% of registering participants in the 0-cent category, 25.1% in the 3-cent category, and 39.2% in the 10-cent category.

Further analysis of the data shows that the average task complexity for the first assignment completed by each participant was lower (158) than the average complexity among all tasks (173). As MTurk presents participants with tasks in random order, a significant deviation from batch average for the first task means that uptake is affected by the upfront complexity. Payment level affected this first-task complexity average with scores for the different payment groups being 139 for 0 cents, 169 for 3 cents and 181 for 10 cents. The for-profit group averaged at 153 and the non-profit group at 164. Figure 4.5 shows how task complexity changed as participants completed more tasks. The expected average task complexity was only achieved after participants had completed 15 assignments, while participants completing many HITs had

		Completed assignments	Mean complexity of completed assignments	Working participants	Uptake	N. American workers	S. Asian workers	Female workers	Assignments/worker	Mean task accuracy
Non-profit	0 cent	98	128.3	32	12%	28%	54%	52%	3.1	0.83
	3 cent	200	173.1	26	31%	35%	51%	56%	7.7	0.83
	10 cent	200	173.1	16	41%	31%	49%	45%	12.5	0.73
For-profit	0 cent	132	147.6	36	14%	31%	53%	56%	3.7	0.71
	3 cent	200	173.1	33	22%	35%	47%	45%	6.1	0.66
	10 cent	200	173.1	15	38%	48%	35%	38%	13.3	0.75

TABLE 4.1: Performance metrics for the six study conditions. Uptake refers to the ratio of qualified participants who chose to complete at least one assignment.

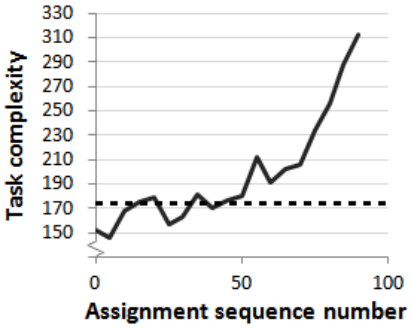

FIGURE 4.5: Average task complexity by assignment sequence number. The dotted line shows the average complexity in the entire workload.

a very high average complexity because they completed difficult HITs that others presumably chose not to work on.

Participants' region of residence also affected performance. As mentioned previously, 42% of participants reported having lived only in South Asia, while 35% had lived only in North America. Yet, 72% of assignments were completed by Asians and 15% by Americans. On average participants from North America completed 2.8 HITs with 89% accuracy and 123.5 mean task complexity, while those from South Asia completed 11.2 hits with 71% accuracy and 172.9 mean task complexity.

The effect that variations in payment had on the number of completed assignments per participant in these two worker groups can be seen in Figure 4.6. A two-way between-groups ANOVA showed a significant main effect of location [$F(1, 115)=10.9$, $p=0.001$] and payment [$F(2, 115)=3.5$, $p=0.034$]. The effect of both of these variables was moderate (eta squared=0.086 and 0.057 respectively). Post-hoc comparisons using the Tukey HSD test indicated significant differences only between means of the 0 and 10 cent groups. While increasing payment levels generally lead to increased work effort for participants both from South Asia and North America, going from a 0 to 3 cent reward appears to have had no effect on Americans.

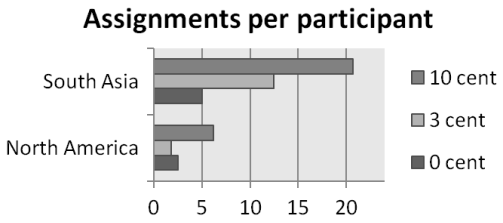

FIGURE 4.6: Breakdown of total work effort (average assignments per participant) by payment level and participant location.

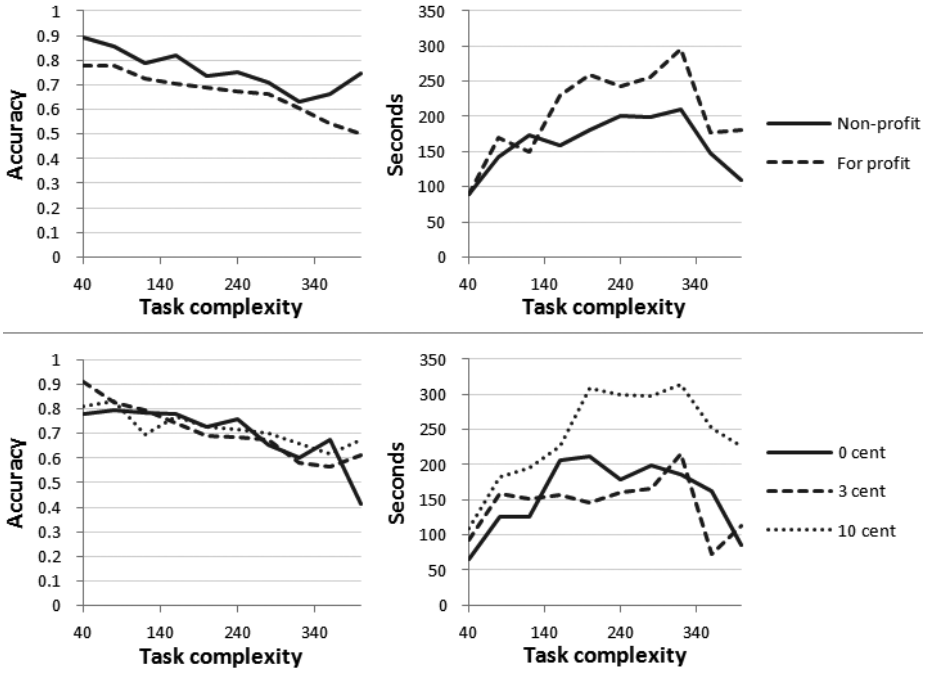

FIGURE 4.7: The effect of variations in task complexity on task accuracy (left) and time spent per task (right).

We also observed that using a non-profit cover story slightly increased uptake and average assignments per participants for Americans and decreased it slightly for Asians, and while these observations are similar to results by Chandler and Kapelner (2013) the effects in our study were not statistically significant. Differences in both work effort and accuracy based on region of residence were clearer than differences based on income.

The time which participants spent on tasks of low complexity was consistent across conditions (Figure 4.7). It then began levelling off around complexity of 100, but peaked at different levels for different conditions. Participants in the for-profit and 10-cent groups spent more time working on complex tasks than participants in the other conditions. Working time decreased significantly for all conditions at the highest complexity levels.

4.6.4 Accuracy

The second metric of interest is work quality, which we quantify as accuracy. An ANOVA showed a significant main effect of "cover story" (non-profit, for-profit) on the accuracy of completed tasks ($F(1,1024)=38.1$, p¡0.0001), while reward had no significant effect. Although we report accuracy scores here based on absolute errors, these errors were almost exclusively underestimates of the true values.

Returning to Figure 4.7, we see the effects on task accuracy and time spent on tasks from increasing levels of task complexity. Accuracy decreased with increasing complexity in all conditions, with participants doing non-profit work being consistently more accurate than for-profit workers. While there appears to be no correlation between time spent on a task and achieved accuracy, task times reported in MTurk are generally not reliable as workers often may have multiple windows and tasks open at once.

Table 4.1 also shows how the mean complexity of completed assignments in the two incomplete 0-cent conditions was lower than in the four completed conditions. This indicates that participants chose to complete only the easy tasks, presumably because the incentives were too small to motivate the effort of working on the most complex tasks. As accuracy decreased with increasing task complexity, this selection effect needs to be accounted for when comparing the mean accuracy between conditions and we thus conclude that participants in the 3-cent non-profit condition produced the most accurate results.

Finally, Figure 4.8 considers how accuracy changed as participants completed more tasks, suggesting that participants under both cover stories performed equally well in their first three assignments. After this, non-profit participants kept gaining in accuracy up to the seventh task, while for-profit participants became less accurate. Beyond this point up to the 25th assignment, both groups became increasingly less accurate, but the performance of non-profit workers decreased slightly slower than others'. The data showed no further accuracy decreases beyond the 25th task, but sample sizes for these levels were limited to only a handful of workers. The decrease in accuracy from increasing numbers of completed tasks cannot be explained by the associated increase in task complexity (Figure 4.5), as the average complexity change for the first 25 tasks was too small. As most participants completed only a few assignments, the number of samples on which the series are based decreases rapidly along the horizontal axis and the increasing variance seen in the graph is to be expected. The sample size was not considered large enough to present similar data across payment groups.

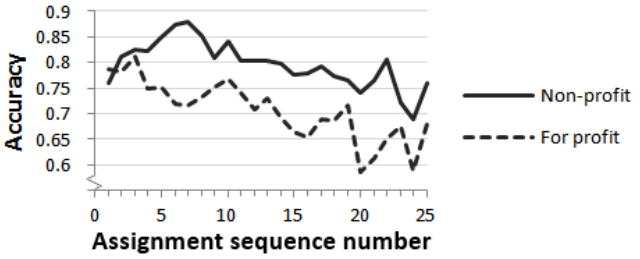

FIGURE 4.8: Mean assignment accuracy by assignment sequence number.

4.7 Discussion

Our motivation for the study was to experimentally assess how workers' performance and effort is affected by varying the levels of intrinsic and extrinsic motivation in a task, as well as examining interaction effects between the two motivational factors. To assess the work completed as part of this study, we measured completion speed and accuracy.

Consistent with prior work, we found that paying people more did not lead to increases in their output accuracy. However, unlike previous work we did find a significant effect of intrinsic motivation on output accuracy: people were more accurate under the non-profit framing than they were under the for-profit framing. Not only was this true for the average task, but also for assignment sequences (Figure 4.8) and for varying levels of task complexity (Figure 4.7). The intrinsic motivation frame did not impact uptake speed; specifically we saw no change in batch completion speed (Figure 4.3), completed tasks per worker and worker uptake (Table 4.1).

We also observed interaction effects between intrinsic and extrinsic motivation, resulting in changes in worker accuracy between conditions that cannot be explained by linear models (Table 4.1). One explanation of these findings consistent with prior theory (Deci 1975) is that intrinsic motivation has a strong positive effect on worker accuracy, but only until the point where extrinsic factors become the main motivator. Further work is needed to explore this and other possibilities.

The hypothesis that increased payment increases work output is confirmed by the data, in full agreement with results from previous studies (Mason and Watts 2009). Higher rewards substantially increased both participant uptake and overall completion rates. This effect may be further strengthened by that MTurk design gives less exposure to low-paying HITs, as it is easy to sort available HITs by reward. Paid participants were also more tolerant to task complexity, as indicated by the average first-task complexities as well as the lower average complexities of completed tasks in the two 0-cent conditions. Participants in the for-profit 10 cent condition in fact exhibited higher-than-average task complexity for their first task. We also find it interesting that although progress in the 0-cent conditions was significantly slower than

in the paying conditions, 12-14% of workers in a task market built around extrinsic motivation were still willing to contribute some work without any form of payment.

Figure 4.6 shows that participants from both South Asia and North America greatly increased their workload once sufficient payment was reached. We note, however, that this sufficient level appears to differ between regions and that Asians were willing to work for less compensation than Americans. The data in Figure 4.5 together with that regional differences were greater than differences between income groups, suggests that both of these rates were high enough to have an effect on Asian workers (from a lower-income society), while Americans (from a higher-income society) and others perceived the 3 cent reward as equal or worse than working without compensation. This finding is in agreement with previous studies showing that if the extrinsic motivation (in this case the reward) is not adequate, performance is likely to suffer (Finin et al. 2010).

4.7.1 Sample bias

Studies of motivation on MTurk, including ours, need to address problems introduced by large differences in sample size for different participants, such as a large number of tasks completed by a small group of participants. This distribution is natural to crowdsourcing markets (and many online communities) in which workers self-select which and how many work items to complete. As our goal is to measure effects of motivation on total work output, our analyses consider the task as our unit of analysis; however, we note that this assigns more weight to people who contribute more work.

An alternative would be to use the worker as the unit of analysis (e.g., calculate means for each worker, followed by taking the means of those means for each condition). In our study, this would have not only biased results towards workers who we know only completed one or two tasks each, but also introduced noise from the great variations in workers' mean task complexity, as well as not being representative of the natural distribution of task uptake.

4.7.2 Strategies and guidelines for crowdsourcing

Below we discuss guidelines suggested by our findings for crowdsourced work.

4.7.2.1 Speeding up progress

The importance of adequate payment on a crowdsourcing market like MTurk is crucial. Not only did higher paying tasks attract workers at a higher rate; those workers also completed more

work once they showed up. This resulted in both higher and more predictable rates of progress. The effect which payment has on progress is simple; higher payment leads to quicker results.

In addition to increased payment, the data shows that quicker results can be achieved by simplifying each work item, which in turn increases uptake of workers.

Our results show no effect of intrinsic motivation on work progress. However, uptake might be improved by highlighting intrinsic value in task captions and summaries, something we could not do due to our study design.

4.7.2.2 Increasing accuracy

Emphasizing the importance of the work (in this case working for a non-profit organization) had a statistically significant and consistent positive effect on quality of answers in the study. Effects were particularly strong at lower payment levels, with differences in accuracy of 12% and 17% for the 0 and 3 cent conditions. These marked differences are surprising given the similarities between the conditions, which both included malaria and the only difference being the company the task was being done for. This difference between conditions was even more conservative than Chandler and Kapelner (2013), who either gave workers a description of purpose or did not.

The results may have application to crowdsourcing charity work, suggesting that lower payment levels may produce higher quality results. It is unlikely that workers actually prefer to work for less money, thus this might suggest that intrinsic value has to be kept larger than extrinsic value for the accuracy benefits to appear.

Although in this study we specifically investigate the non-profit/for-profit distinction as our method of investigating intrinsic motivation, there are a number of other possible ways for affecting intrinsic motivation as well. Future work investigating factors such as social identity, goal setting, and feedback could all be profitable directions (Cosley et al. 2005; Beenen et al. 2004).

4.7.3 Demographic considerations

Although most work in this study was performed by participants from Asia, people from North America were on average more accurate but less tolerant to high task complexity. Such regional differences are worth keeping in mind for a number of reasons. Americans are a large group; a third of the workforce in our study and 40% of site visitors according to statistics from Alexa (www.alexa.com). In addition, nationality is one of the few built in ways of restricting access to work that MTurk supports, without creation of additional qualification tasks. Our data

does however suggest that excluding Asian workers is likely to have severe impacts on work completion rates, in particular if payment is kept at a level which is perceived by Americans as low.

While 158 participants completed work in this study, only nine completed 30 or more assignments and together account for half the total output. Designing tasks that attract these workers may have significant effects on work completion rates and their demographics are therefore worth mentioning. All carried bachelor's degrees or higher and all but one lived in South Asia. Six were male and ages were equally distributed between 18 and 44. Most reported spending more than six hours per week working on MTurk and yearly incomes were generally below $5,000. The participants were equally distributed between the cover stories, but favoured higher paying tasks. The highest work output (95 assignments) was by an Asian woman, 35-44 years old with a bachelor's degree and with a yearly income between $20,000 and $60,000.

4.8 Conclusion

This study has shown that intrinsic motivators significantly improve work accuracy, especially when extrinsic motivation is low. It is encouraging to find that highly intrinsically motivated workers are likely to provide high-quality work, as quality control mechanisms available in crowdsourcing mainly depend on collecting and aggregating redundant input from multiple workers. By its very nature, a scheme that depends on redundant completion of tasks wastes precious work effort, and high quality work output can permit less strict quality control, as well as task designs for which it may be difficult to aggregate the answers of multiple workers.

Furthermore, the interaction between intrinsic and extrinsic motivators appears to be such that workers provide highest quality results when intrinsic motivators dominate over extrinsic motivators. Once extrinsic motivation takes over, accuracy converges to levels that are independent of the work's intrinsic value. The implication of these findings for the design of systems to be used in disaster response is that if the accuracy-related benefits of the work are to be maintained, payment must be kept sufficiently low. However, future work may be needed to investigate to what extent this finding is valid also outside of for-pay task markets.

We also find, consistent with prior work, that increasing levels of payment increases work output regardless of intrinsic value. In addition, payment on for-pay task markets needs to be kept competitively high if the tasks are to attract workers, and if those workers are to maintain an acceptably high rate of output. This directly conflicts with the suitable strategies for maximizing work quality, thus while crowdsourcing markets can in theory provide access to a vast pool of workers, a system that integrates workers from a for-pay market is likely to generate high

operating costs. Furthermore, the quality of work produced by paid workers is likely to be lower than that of volunteer workers.

The study design does not permit us to pinpoint the reason why payment is such a critical factor in the market, but we hypothesize that the platform's user interface design is a major factor in addition to motivational aspects. Mechanical Turk assists workers in finding tasks that maximize their income, thus low-paying tasks are less likely to show up in searches and listings.

Because intrinsic value cannot serve as a substitute for payment in for-pay crowdsourcing markets, these markets are expected to have limited use in continuous processing of large information volumes during humanitarian disaster response. This is true in particular in settings where processing capacity needs to be sustained over longer periods of time resulting in prohibitively high operating costs, such as in monitoring of conflicts, pandemics or the recovery stage after major natural disasters. The information processing techniques proposed in later chapters are found to perform best in these settings, and for this reason future chapters turn to other methods of recruiting workers.

For-pay markets may however have a role to play in the early hours following a rapid onset disaster, before sufficient numbers of volunteer worker have been recruited to transition to a no-pay approach. As workers on task markets can be reached computationally, a system could be up and running within minutes as long as sufficient funding is available. If future research pursues this direction, significant care needs to be taken to not cause long-term persistent negative effects on work quality beyond the point when money is no longer offered, similar to what occurred in an experiment by Gneezy and Rustichini (2000a).

Chapter 5

Exploratory analysis of clustered microblog feeds

Previous chapters have identified several ways in which citizen reports posted on social media during disasters differ from document corpuses used in traditional information retrieval and information extraction research. The properties of social media content include low signal to noise ratio, very short document length, great variety in grammar and language, that content spreads largely through duplication (rather than reference), and that messages arrive as a stream of up to millions of items per day.

Because of these differences between social media and traditional document corpuses, new efficient solutions are sought that can perform de-duplication, grouping and summarization of related content. In addition to reducing redundancy, improvements in signal-to-noise ratio are needed to handle information volumes that can be expected to almost always exceed that which can be directly processed by humans. Finding automated solutions for these processes would directly reduce wasteful human processing of redundant information that in no way contributes to situational awareness, both among crowdsourcing workers and end-user information consumers.

5.1 Study goals and methodology

These desired properties can in theory be achieved through clustering, a common data mining technique in which some similarity metric is used to group together similar items into clusters, and to separate dissimilar items into different clusters. Aggregate properties of the clusters, such as the number of items, or mean or variance of some item property, can be used to rank different clusters, and summarization algorithms can be applied to condense the cluster content into digestable bits of information. In an online setting, new or rapidly growing clusters can be highlighted as breaking news.

Under the working hypothesis that text-based clustering will let users consume Twitter activity on the basis of distinct pieces of information, rather than individual tweets, a software prototype was implemented that performs online clustering of Twitter content and allows a user to explore the clusters and their contents.

This chapter presents an exploratory analysis of the clustered output, collected during five large-scale events. The purpose of the analysis is to provide early indications of the degree to which a state of the art clustering algorithm can be used to cope with the many challenges imposed by processing of citizen reports collected during ongoing disasters. Content was explored using the custom user interface of the prototype system, as well as through numerous database queries, for instance to assess to what extent reports relating to a single event was split across different clusters and to find out if unrelated messages had been incorrectly associated with large clusters.

The analysis looks at the degree of redundancy in Twitter message streams, and considers different computationally feasible ranking metrics in terms of newsworthiness and resistance to spam content. It was first published as a workshop paper (Rogstadius et al. 2011a), but has been partially rewritten to fit the format of the thesis.

5.2 Stream clustering using Locality Sensitive Hashing

In practice, very few algorithms have been proposed for online language-independent clustering of very short messages, which operate in linear time and space with regards to the number of documents. One such algorithm is Locality Sensitive Hashing (LSH) using cosine similarity of bags of words to assess the relatedness of documents (Charikar 2002). This version of LSH is in principle an efficient online implementation of k-nearest neighbour (kNN) clustering, using sets of random hyperplanes as hash functions to sort a stream of feature vectors into hash bins. The probability of two documents being similar according to the cosine metric is then proportional to the probability of the two documents ending up in the same hash bin. By using multiple hash tables, the algorithm identifies probabilistic near neighbours as documents that occur in the same bins across multiple hash tables. The real similarity score is then calculated between the new document and the top set of probabilistic nearest neighbours, to find the N nearest neighbours, and the item is assigned to the majority cluster within this set. An item is considered a new cluster if it has no neighbours within a threshold distance.

It is worth mentioning that for very short documents, such as tweets, the cosine similarity metric will give a zero similarity score for most random documents pairs. With small feature sets (word), drawn from a large feature space (a language), the similarity metric will only work locally. LSH-based clustering overcomes this by being an agglomerative algorithm, in which

dissimilar items a and b are grouped if they are bridged by a third item c, which is similar to both a and b.

An implementation of this algorithm extended with a separate buffer for recent messages has previously been applied for first story detection in a Twitter corpus of 160 million messages (Petrović, Osborne, and Lavrenko 2010). This study looked at metrics for ranking tweet clusters to maximize the ratio of top clusters that were classified as news, but only evaluated clustering performance itself on a separate newswire corpus. The algorithm's true performance in terms of precision, recall and similar metrics is thus unknown for Twitter content, and will likely remain so; Petrović et al. note that this would require manual annotation of millions of items.

In addition to the uncertain clustering performance, the algorithm has many tuning parameters that need to be set manually, and which if poorly configured can at worst cause all items to become one cluster, or each item to become a cluster of its own. In addition, work on EMM NewsBrief (Best et al. 2005) showed that it is desirable for clustering algorithms applied to news information to support branching and merging of clusters to capture how events unfold over time. These features are not supported in the LSH algorithm and it is not trivial how to modify it to introduce this support.

Despite its limitations, LSH remains one of the best options currently known to be available for computationally feasible online language-independent clustering of high-volume streams of short messages.

5.3 Prototype system

To better understand the effects of applying online clustering to citizen communication during disasters, a prototype system was implemented, consisting of a processing pipeline for clustering, and a user interface for exploring the clustered messages. In short, the system enables administrators to define high-level topics, and over time a number of clusters, referred to as stories, emerge within each topic, driven by activity on Twitter. The central use of stories is the primary difference between this prototype system and previous systems that visualize Twitter activity, such as Twitris (Jadhav et al. 2010), Twitcident (Abel et al. 2012) and Eddi (Bernstein et al. 2010).

5.3.1 Processing pipeline

Using this prototype system, an administrator can define major high-level events (e.g. an earthquake) to track by describing them as sets of weighted keywords. Such a set of keywords will be referred to as a topic. An admin-defined subset of the topic keywords are tracked using

Twitter's streaming API, resulting in an incoming stream of tweets; each containing at least one of the tracked keywords and therefore considered potentially relevant to the tracked topics. The tweets are then split into words, stemmed, and terms are weighted using estimated inverse document frequencies (see below) to generate a term-weighted word vector for each tweet. The cosine distance between the tweet vector and each admin-defined topic vector is then calculated, to determine the tweet's general "appropriateness" for each topic. If the similarity is within a threshold and if the tweet contains enough information (as estimated by the length of its word vector), it is assigned to its most similar topic, else it is discarded.

Once assigned to a topic, tweets are clustered using a custom implementation of LSH, based on the suggestions of Petrović, Osborne, and Lavrenko (2010). While Petrović et al. used an initial computation pass to calculate global word statistics (inverse document frequencies) in their offline corpus, such a global pass is not possible in a true online setting. Word frequencies cannot be assumed to be constant over time, e.g. due to local changes in the tracked high-level event and global activity in different time zones. The algorithm was therefore extended in two important ways.

First, word statistics are collected based on both the filtered stream and Twitter's sample stream, a 1% sample of all posted tweets. The word distributions are then approximated with a simple IIR filter with exponential decay and a four-day half-time. Words that have a global frequency greater than 90% of the maximum frequency are labelled as stop words unless they are tracking keywords. Words that have been seen less than three times are ignored. The second extension is to replace the oldest hash function hourly with a new function created from the current dictionary. The new hash table is then populated with the items from the removed table.

Each cluster is referred to as a story, with each story containing tweets that are closely similar to each other in terms of their word distributions. The system does not distinguish between source tweets and retweets, and in fact, due to design of Twitter's API, it is possible that the system never receives the original tweet or that it arrives after a retweet. The system does keep track of the time when the first tweet in a story was detected, using this as an estimated timestamp for the information contained in the story. Each story is assigned a title using the text in the most representative (centroid) tweet within the story. As the system uses an agglomerative clustering method, the centroid is calculated as the tweet to which the highest number of other tweets was classified as similar.

5.3.2 User interface

Users of the prototype system can access the information through a web interface (Figure 5.1). On the left is a list of currently tracked topics and by clicking on a topic the interface becomes

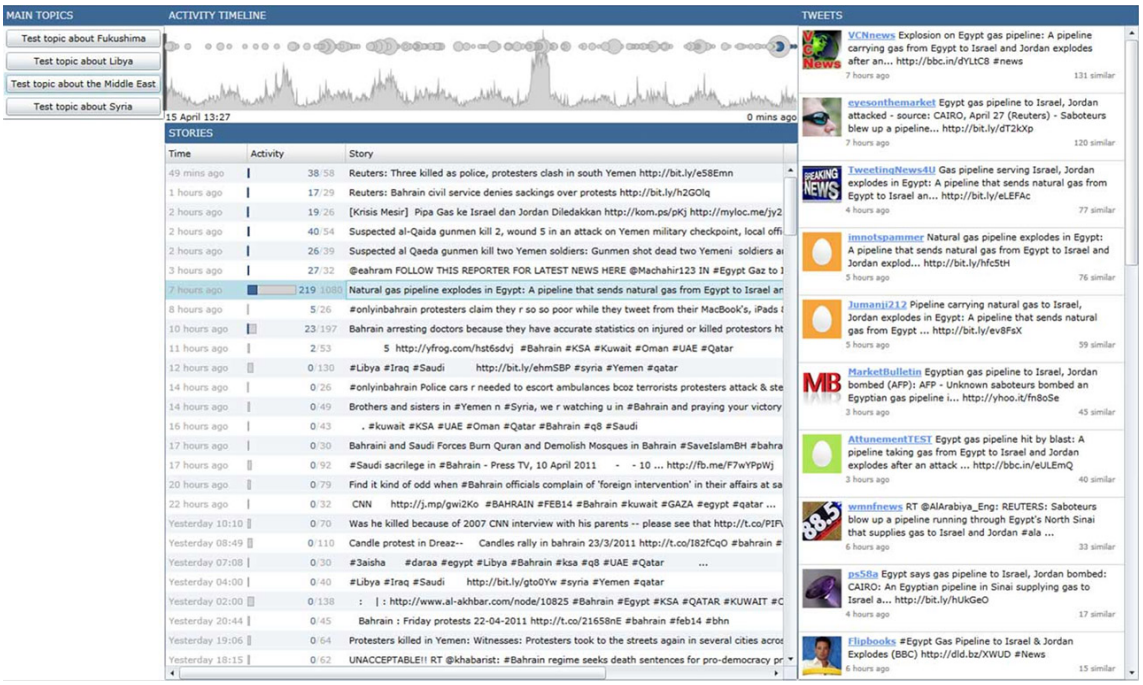

FIGURE 5.1: The web interface of the prototype system.

populated with information related to that topic. Along the top is a graph presenting overall tweet activity (total number of tweets) over time and below that is a reverse-chronological list of detected stories. Stories are also overlaid as circles on the activity graph, with circle size corresponding to the number of tweets contained in the story. Dark blue segments on story bars and circles represent tweet volume over the past two hours.

By clicking on a story, either in the list or on the timeline, a viewer can bring up the message variations that together make up the story. Here, messages are grouped by the leading 20 characters in the message and sorted by descending group size, which effectively collapses retweets lacking added comments into a single item.

Stories containing few tweets are presented when those stories have very recently been detected, but as stories age, increasingly large stories get removed from the interface. This design decision was made to avoid cluttering while still maintaining a long-term summary of key events around the topic.

5.4 Data collection

The prototype was used to track five different large-scale events during the spring of 2011 for periods of four to eight weeks each and totalling 300,000 to 500,000 topic-related tweets per event, posted by 52,000 to 142,000 users. The tracked events were the nuclear disaster at the Fukushima plant in Japan; the civil war following public protests in Libya; political protests in

Syria; protests in various countries in the Middle East (multiple countries together), and the final and semi-finals of the Champions League.

5.5 Results

5.5.1 Sub-event detection

The time sorted list of clusters detected by the system provides a narrative of recent sub-events. The time stamped list in Table 5.1 is an extract of cluster titles displayed by the system on April 22, 2011, a day of widespread protests and civil unrest during the early build-up to the Syrian civil war. Notable is that several messages contain rich information such as references to locations of demonstrations with estimations of number of people taking part, demands by protesters, deployment of anti-riot police, and finally reports of violence.

5.5.2 Story composition

Based on manual inspection of cluster composition for each of the tracked events, the clustering algorithm does appear to offer high precision, by managing to bring together content with high semantic similarity, without significantly including other content that happens to be textually but not semantically similar. An example of such semantic similarity is the two messages "#Syria; 60,000 protesters are gathering in Hama!" and "Thousands of protesters in the center of #Hama right now #Syria". Stories also contain many retweets, which, in addition to repeating a message, were observed to often contain added information or comments from the retweeting user[1]. Therefore, not only the first tweet in a story is of interest from an information point of view, as many aspects of the event can be captured in later tweets.

However, the implemented algorithm suffers significantly from low recall, in the sense that reports mentioning the same event are often split across several clusters. This is important, if each cluster is to be treated in a system as a new piece of information. Redundant clusters can likely be explained by two properties of the LSH algorithm. First, the bag-of-word approach only captures similarity between messages where their sets of words directly overlap. Use of synonyms and other semantically related terms is not captured. Second, new clusters are formed based only on the uniqueness of the first tweet in a story. If several independent reports are posted early after an event that differ significantly in their word usage, these will be considered as separate clusters regardless of if later messages effectively bridge the gap between the first

[1]This research was conducted in 2011. As of 2013, Twitter has refined its retweeting mechanism and non-verbatim retweets are now less common. Textually different tweets capturing complementary views of the same story remain common.

07:06	Syria tense ahead of new protests: Syria tightens security ahead of what protesters say will be the biggest rall... http://bbc.in/gC91eg
11:37	Heavy secret police checkpoints at major entry points into #Damascus, #Syria. People report being stopped 3-5 times driv ...
12:03	#Syria churches cancel Good Friday street processions ahead of expected protests http://aje.me/h4Jm4i
13:12	Anti-riot police have been depolyed around mosque's in #Damascus, #Syria. Government is clearly taking today VERY seriously.
13:13	More than 30,000 protest in Douma now. That's it, I am moving in to Doumaaaaa! #Syria #fb
13:32	Pro-democracy protest in historical Midan district in central #Damascus: witness #Alarabiya #Syria
13:35	More than 7000 protesters r now in #DairZour #Syria
13:37	Around 10 thousands protesters are now in #tal #Syria
13:42	Huge demonstration in Al-Zabadani following Friday prayers in Al-Jisr mosque. #Syria #fb
13:54	Protesters in Latakya attacked with tasers by regime's thugs and forces. #Syria #fb
14:08	Protests in, #Damascus, #Homs, #Deraa, #Banias, #Latakia, #Hasaka so far confirmed. #Syria
14:15	SYRIA: As protests get underway, activist says there's no going back [LAT] http://lat.ms/hH9Sp9 #Syria
14:29	I confirm live gunfire in #Homs, different parts, to disperse thousands of protesters. #Syria #March15
14:33	Syria: 'Troops Fire Teargas At Protesters' http://t.co/A7pScL3
14:36	First protester killed in #Homs, #Syria today.
14:38	Reports coming in of live rounds being used in #Hama, #Syria against anti-Assad protesters.
15:12	Sign from a protest: "We want liberty, not an amendment to our enslavement" #Syria
15:32	http://bit.ly/dU2nLa Video shows protester killed by security forces in Qaboun suburb of #Damascus #Syria
15:33	Totals thus far: At least 7 dead in village near #Deraa. At least 1 killed in #Homs. At least 2 killed in #Duma. #Syria
16:03	Far too many tweets to be able to RT. #Syria is on fire today. Demonstrations everywhere that are violently repressed w. heav...

TABLE 5.1: An extract of timestamped cluster titles produced by the prototype system on April 22, 2011.

reports. The next chapter will return to the issue of recall, proposing strategies for handling such undesired branching.

Unlike Petrović, Osborne, and Lavrenko (2010) who looked at first story detection across all Twitter activity, the prototype system uses Twitter's filtered stream to only collect messages containing any of a set of predefined keywords. These preliminary results suggest that the clustering algorithm looks promising also for corpuses with less textual variance.

5.5.3 Degree of repetition of information in the dataset

Out of all the tweets we processed, 29-47% of tweets belonging to each topic were retweets. For the purposes of this study, a retweet is defined as a tweet containing the standard notation "RT @username", but as users can use other notations, actual numbers are likely higher. Furthermore, the system's processing pipeline classified as many as 60-75% of all incoming tweets (varying by topic) as being highly similar to at least one other tweet. The largest single cluster detected by the system (a NATO airstrike that killed Col. Gaddafi's son Saif al-Arab and three of Gaddafi's grandsons) contained 10920 tweets during 24 hours, plus several smaller clusters on the same story that were incorrectly split off by the system. These numbers together strongly suggest that clustering can be a useful first step in coping with information overload during large scale events, but that the recall of the selected clustering algorithm may need to be improved.

5.5.4 Repetition as an indicator of importance

A system that attempts to extract salient information from a stream needs metrics for information prioritization and for noise removal. However, different metrics are suitable for different user groups and we see three general types of information that such a system needs to identify and work with.

- Information that enables members of the public to follow events and crises that they are not directly affected by. This includes links to news articles that summarize recent changes and real-time mentions of highly influential events.
- Information that helps people directly involved in the crisis to make decisions, e.g. victims and emergency responders. This can be, for instance, locations where help is provided after an earthquake, but also higher level information that has implications for the days to come, such as agreements made between parties involved in a conflict.
- Information too fine-grained to be picked up by more than a few individuals on the tracked social media platform, but which when combined with other information pieces provides the necessary input data to a detailed computationally generated event model.

5.5.4.1 Information for the general public

Previous work by Starbird and Palen (2010) has suggested that a large number of retweets is an indicator of high-level information that appeals to a broad audience during a crisis. Our informal evaluation confirms that stories containing hundreds or thousands of tweets are indeed of greater interest to the general public than those containing only a few messages, though the

exact size of one story relative to another appears to be quite random. This randomness could be due to limitations in our clustering algorithm or due to that we disagree with the general public on what is interesting. It could also indicate that Twitter's information propagation structure introduces randomness in ways similar to what Salganik, Dodds, and Watts (2006) observed in the popularity of songs in an artificial music market. Though we lack a controlled study, our impression is that like in their work, stories containing information that is of little public interest are always small and stories of great public interest are always large, but the amount of Twitter activity around a story of medium-level public interest cannot be predicted well based on information content alone.

Most top stories in our dataset related to the conflict topics contained links, while the top stories related to the Champions League generally did not.

5.5.4.2 Information for domain experts

Information of value primarily to a local audience (particularly useful for crisis intervention) is not well captured by measuring the overall Twitter activity. Starbird and Palen (2010) suggest that a better metric is retweeting of information among users who are local to the event, but to get this information, they relied on manually examining the message history of each Twitter account to determine who the local users are. For our streaming scenario we have instead looked at more computationally feasible metrics, of which the most promising appears to be to keep track of the number of topic-related tweets that have been posted by each user. The most frequently seen users are then treated as a set of domain experts, and message clusters are ranked by the number of expert users whose tweets are found within each cluster. This metric also acknowledges the fact that not all domain experts may be physically located on-site.

It is challenging to assess *how* the content posted by these "expert users" differs from that of other users, but we can at least look at *how* much the content differs. Spearman's rank correlation scores between the total number of users contributing to a story and the number of expert users who contributed to the story is fairly low, around 0.6 for all topics. Furthermore, during the time of this study, protests in Syria were taking place primarily on Fridays and total tweet counts for the Syria topic was on average twice as high on Fridays as during other days. Peak activity during Fridays was around ten times higher than average daily activity. To see if this volume difference reflected a difference in type of content, we looked at the weekday of the detection of each story as well as which users participated in sharing of the story. We found that only 11 of the top 50 stories among the general public discussed events on Fridays, while for domain experts this number was 19.

Both these metrics hint that the information shared by "domain experts" is indeed of a notably different nature than that shared by the general public. This further suggests that content ranking using a large subset of accounts somehow detected to be affiliated with a class of reports may be a feasible approach to highlight content of interest to a specific target audience. Alternatively, instead of using purely computational techniques to identify domain expert tweeters, a crowdsourcing approach could be used to classify users based on their tweet history.

5.5.4.3 Information for event modelling

The system cannot yet detect stories or individual tweets that contain who-what-when statements (e.g *"The fourth division in the army are now bombing Alrastan from four sides with T-72 tanks #Homs #Syria"*) that could contribute to a larger event model. However, aside from the text processing issues involved in parsing text that often contains informal grammar and spelling, we have found a few Twitter users who invest significant effort on posting such updates. Often these posts are so fine-grained that they are never retweeted. Identifying such users and tracking their tweets may be a good future approach to collect data with minimal noise.

5.5.5 Coping with spam

We have observed that a common spamming approach on Twitter is to post messages containing a short message (e.g. *"Cool pictures"*), a large set of popular keywords in random order, together with a shortened URL. In relation to the political protests in the Middle East, we also observed what appeared to be an attempt at a form of denial-of-service attack, where a large number of messages containing only popular keywords would be posted from multiple accounts. During some time periods, so many messages were posted that it became difficult to find legitimate posts containing these keywords using the search provided by the official Twitter website.

By merging similar tweets into clusters, the prototype system represented this flood of spam messages as one or a few large stories, sometimes containing an order of magnitude greater number of tweets than other top clusters. While the spam was noticeable, the overview provided by the system did not suffer from problems of information occlusion.

A second observed type of spam took the form of one or multiple users that posted the same story up to hundreds of times. This spamming strategy is well dealt with by using number of unique users as an importance metric for stories, but not by using total number of tweets. Pearson correlation between total number of user mentioning a story and the total number of tweets in the story in our dataset always surpasses 0.95, thus in the non-spam case the numbers are equivalent.

5.6 Conclusion

This initial exploratory analysis suggests that cluster precision will not be a significant issue when applying agglomerative text-based clustering to Twitter content. The LSH algorithm (Petrović, Osborne, and Lavrenko 2010) effectively brings together messages that are semantically related. Furthermore, the algorithm works reasonably well for new event detection (separating messages that mention new versus old information), but many false positives are detected as cluster recall is an issue that that needs to be substantially improved beyond what is achievable using the previously published algorithm.

In agreement with previous research, cluster size alone appears to be a poor metric of how interesting a piece of information is to any specific stakeholder group. However, large subsets of users can be identified computationally to be used to identify content of interest to specific groups of users.

The analysis also identified two spamming strategies, and observed that clusters are better ranked by their numbers of contributing users than the number of messages. Clustering by itself also helps cope with spamming strategies that attempt to flood legitimate discussion with nonsense content.

Based on these conclusions, clustering is considered to be a highly promising approach for addressing the primary challenges that cause information overload when trying to monitor social media communication during large-scale events. The next chapter continues the development of a social media monitoring tool and evaluates its use in a real-world conflict analysis setting.

Chapter 6

Utility and scalability of hybrid processing

The exploratory analysis in the previous chapter concluded that clustering of social media feeds can be a feasible approach to identify unique pieces of information in a torrent of short and highly redundant messages. Once clustered, ranking metrics exist that show early indications of being able to separate signal from noise. However, this information collection methodology needs to be assessed more thoroughly with regards to its utility for increasing situational awareness among stakeholders in humanitarian disasters.

This chapter presents a field evaluation of CrisisTracker, a system supporting real-time social media curation, using a hybrid approach that integrates machine-based pre-processing with human computation. The study was first published in the IBM Journal of Research and Development (Rogstadius et al. 2013b) and has been edited to fit the structure of the thesis.

6.1 Study goals

From previous work, social media is known to contain extensive information relevant to various stakeholders in disasters. Moreover, clustering of social media, using the techniques introduced in the previous chapter, can be an effective way to detect breaking news and to summarize content. This study will investigate if the proposed information processing techniques effectively highlight social media content that corresponds to the information needs of disaster response professionals. In addition to the relevance of information, the study will assess if breaking news is detected early enough for the approach to be competitive with other methods of information collection.

This study thus aims to go beyond purely theoretical assessment to understand the contributions of introducing a social media summarization tool into real-world humanitarian disaster

information management. To what extent can a system built using the proposed techniques contribute information that is novel or complementary to that received through other channels already in use? Does the use of CrisisTracker improve situational awareness, and if so, to which of the situational awareness components (perception, comprehension, projection and prediction) does it contribute?

The study also quantitatively measures the extent to which machine-based report collection and clustering can improve the scalability of crowdsourced human computation, in a system where workers are given annotation tasks similar to those in the popular Ushahidi system. Based on the observations, suitable strategies for worker management are discussed.

Finally, open-ended feedback is collected from expert users and crowd workers to find areas of improvement, and to understand how the proposed system may alter existing work practice.

6.2 System design

The prototype system introduced in the previous chapter performed machine-based data collection, report de-duplication and breaking news detection. It however lacked any functionality for meta-data extraction, which earlier chapters explained is needed to filter and aggregate information into a format that matches stakeholder needs. For disaster-related social media content, the prevalent solution to meta-data extraction to date is crowdsourced human-based computation, as implemented in the Ushahidi system.

This first iteration of the CrisisTracker system combines these two approaches; it automatically collects and ranks news stories related to an ongoing disaster, and lets a human crowd collaborate to curate the stories through meta-data annotations and cluster refinement. An overview of the processing pipeline is shown in Figure 6.1. This section describes each system module in more detail.

6.2.1 Machine-based data collection

CrisisTracker leverages the existing user base of the Twitter microblogging service to gather information about ongoing events. Tweets are collected through Twitter's streaming API, which allows a system administrator to define filters in the form of words, geographic bounding boxes and user accounts for which all matching new tweets will be returned as a stream. Twitter's API returns messages tagged with any geographic region that partially overlaps with specified bounding boxes. As regions can be very large, the system performs an additional filtering pass to discard such messages and keep only those explicitly tagged with a coordinate within bounding boxes of interest (assuming no other filter is matched). The tracking filters are constrained

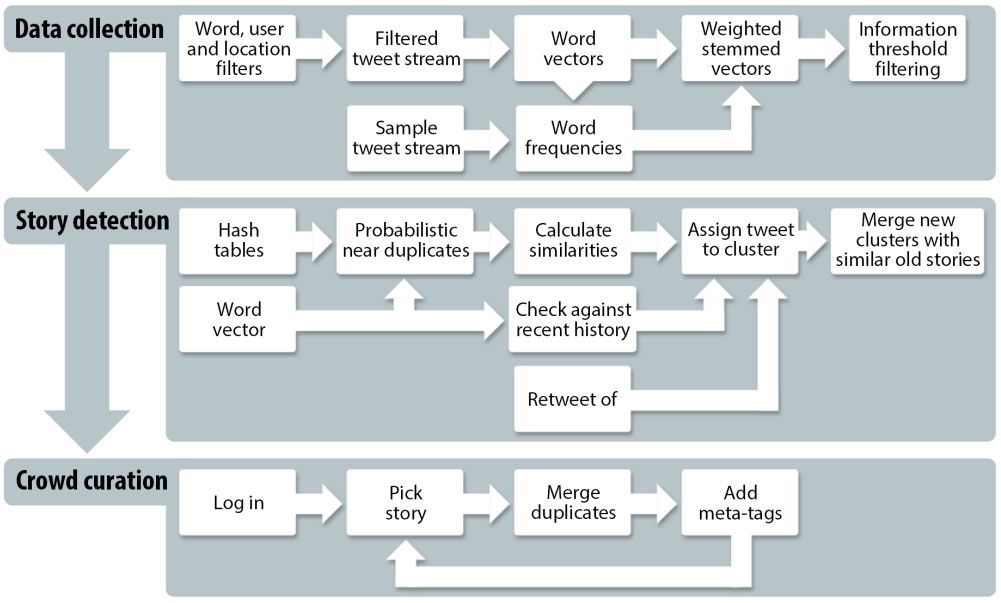

FIGURE 6.1: CrisisTracker's information processing pipeline, consisting of three modules – data collection, story detection and crowd curation. Word frequencies from the data collection module are shared across all modules, which is not shown in the figure.

by the API and the system cannot yet suggest keywords or user accounts to track. Generally around 1% of all tweets are geotagged and it is infeasible to manually select user accounts to cover truly large-scale events, thus good keyword filters are the primary way to obtain high information recall in the system.

CrisisTracker also discards messages having fewer than two words after stop word removal and a very low sum of global word weights (approximated inverse document frequencies). In practice discarded messages are mostly limited to short geotagged messages without any context (e.g. *"@username Thanks!"*).

Selecting a good set of tracking filters is a very important part of setting up the system, as this configuration directly influences what content enters the system. A poor selection of filters means that no useful reports will be picked up, and no amount of curation will be able to improve the utility of the corpus. Depending on the nature of the tracked disaster, the set of filters may need to be modified over time, and a natural system extension would thus be to include an algorithm that can continuously identify high signal-to-noise terms in the feed.

6.2.2 Machine-based de-duplication and breaking news detection

Incoming tweets are compared to previously collected tweets using a bag-of-words approach, meaning that the textual content of tweets is treated as a weighted set of unsorted words, and that tweets are more similar the more words they have in common. A cosine similarity metric is then used to group together messages that are highly similar. Clustering is performed using Locality Sensitive Hashing, a probabilistic technique using hash functions that quickly detects near-duplicates in a stream of feature vectors. For a more in-depth discussion of the strengths and weaknesses of this algorithm, see section 5.2. The time range within which near duplicates can be detected for any incoming message depends on the rate at which similar messages have been received, and interested readers are referred to (Petrović, Osborne, and Lavrenko 2010) for details.

The exploratory analysis of cluster quality in the previous chapter concluded that the original thread-based clustering algorithm from (Petrović, Osborne, and Lavrenko 2010) offers high precision, such that clusters typically contain only highly similar tweets. However, the algorithm's recall was relatively low, such that the set of tweets that discuss a particular event was often split across several clusters.

Therefore, CrisisTracker uses an additional second-order clustering pass, in which each new cluster is compared against those clusters that received the most new items in the past four hours. For this comparison, top clusters are represented by their centroid, the truncated and weighted aggregate word vector of their child items. New clusters are merged with a past cluster if their cosine similarity is above a high threshold, or if they are significantly more similar to one cluster than to all others. Such a cluster of clusters is referred to as a story and as the next section explains, this method also enables human intervention in the clustering process. Initial informal evaluation suggests that our approach greatly improves recall without substantially reducing precision, but accurate measurement of cluster recall for Twitter-scale corpuses is a research problem in itself and not the focus of this chapter. We therefore leave the specific cluster evaluation for future work and instead focus on evaluating CrisisTracker's general ability to improve situation awareness and support decision making.

A limitation of any bag-of-words based event detection technique is that clusters do not necessarily correspond to events, as tweets can potentially have high textual similarity and be grouped together without discussing the same event. This has been observed to cause issues with Twitter accounts that publish automated sensor-based feeds consisting of regular updates about weather or earthquakes, where each message follows a standardized template. The system will group all these messages together, as they differ only in one or a few words, despite that they refer to completely different events.

6.2.3 Meta-data extraction through crowd curation

One purpose of clustering the tweet stream into stories is to facilitate crowd curation. Much of the stream content consists of messages that repeat information already seen in earlier messages, and de-duplication eliminates, or at least greatly reduces, the amount of redundant work that needs to be performed. Clustering also enables size-based ranking of stories, and brings together messages that describe the same event, but which contain different details necessary for piecing together a complete narrative.

Search and filtering requires meta-data for stories. Some of this meta-data is extracted automatically, i.e. the time of the event (estimated as the timestamp of first tweet), keywords, popular versions of the report, and the number of unique users who mention the story (it's "size"). A story's size over the past four hours represents how "active" it is. Story size enables CrisisTracker to estimate how important the message is to the community that has shared it (Starbird et al. 2010; Rogstadius et al. 2011a). Based on the findings in the previous chapter, users of the system can rank stories by their size among all Twitter users, or among the 5000 users most frequently tweeting about the disaster. Based on subjective experience, the top 5000 option better brings out stories with detailed incremental updates to the situation, while the full rank more frequently includes summary articles, jokes and opinions. Once meta-data is assigned to a story, it also covers future tweets that are classified to belong to the same story.

Human workers in the system, referred to here as "curators", can select stories not yet annotated with meta-data from a list sorted by activity within the past four hours. The first curation step is to further improve the clustering, by optionally merging the story with others in a list of possible duplicate stories that the system has detected are textually similar, but fall below the threshold for automated merging. Miss-classified tweets can also be removed from the story. The curator then annotates the story with geographic location, deployment-specific report categories (e.g. infrastructure damage or violence) and named entities using the interface in (Figure 6.3). Stories deemed irrelevant (e.g. a food recipe named after a location) can be hidden, which prevents them from showing up in search results.

Curators identify themselves in a system through their Twitter account. As demonstrated by Standby Task Force deployments using the Ushahidi system, volunteer curators can be recruited globally, or from within the affected community if post-disaster infrastructure permits.

6.2.4 Information consumption

Disaster responders and others interested in consuming the collected information can filter stories by time, location, report category and named entities. This meta-data structure was selected based on information available at the time, but improvements could be made based on

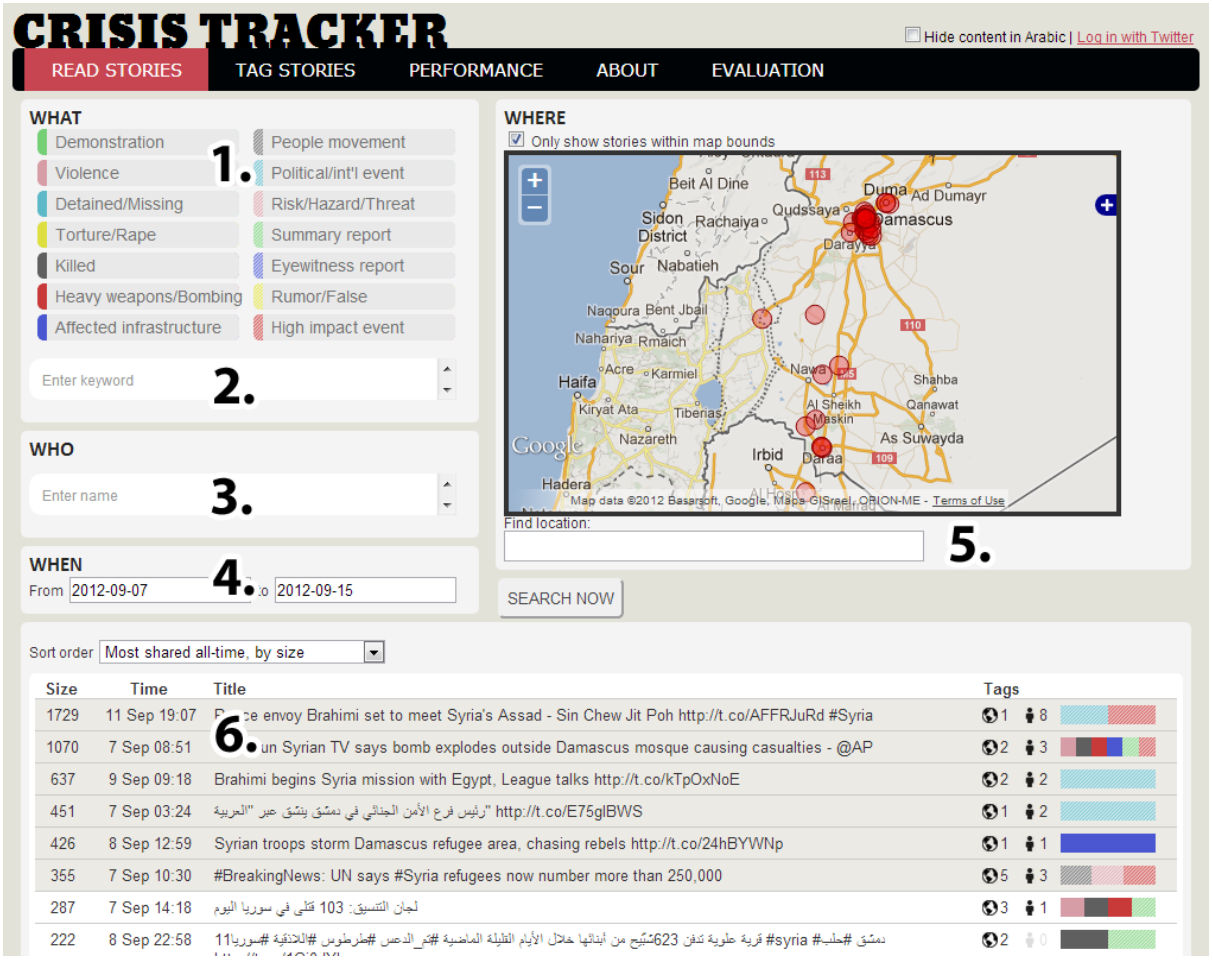

FIGURE 6.2: User interface for exploring stories, with filters for category (1), keywords (2), named entities (3), time (4) and location (5), with matching stories below (6).

the documented information needs presented in section 2.4.3.2. Figures 6.2 and 6.3 present the interfaces for exploring stories and for reading and curating a single story.

6.2.5 Storage

All collected content is stored in a relational MySQL database. As the system would quickly run out of storage space if all content was kept, increasingly larger stories and all their content are deleted with increasing age, unless they have been tagged by a human. Stories consisting of a single tweet are kept for approximately one day.

6.3 Evaluation methodology

To evaluate how CrisisTracker supports its users, a field trial was conducted focusing on the 2012 civil war in Syria. Syria has a population of 21 million, high Twitter adoption and users

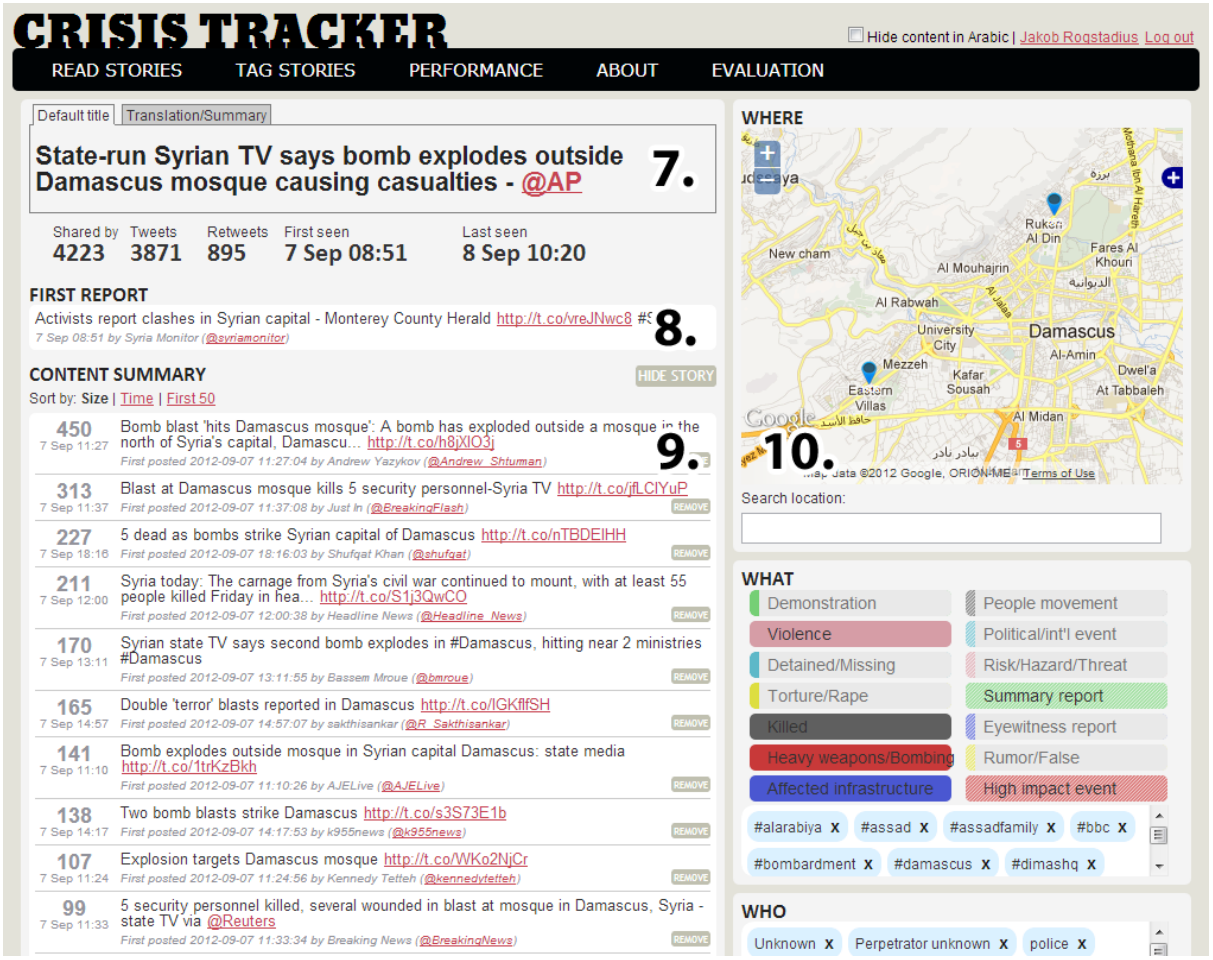

FIGURE 6.3: A single story, with title (7), first tweet (8), grouped alternate versions (9) and human-curated tags (10).

post tweet in both English and Arabic. The conflict thus includes many of the characteristic challenges of social media information management during international disasters.

We recruited 44 unpaid volunteer curators to participate in the eight-day field trial. Half participated for a single day only, three participated every day and on average 13 unique participants were active each day. Volunteers were recruited through the Standby Task Force or independently. Approximately two thirds had previous experience from online disaster volunteering and one third regularly participated in onsite disaster response. Curators were provided with detailed instructions on how to use the system to organize and annotate stories, and could interact with each other and with the researcher through a dedicated persistent text chat.

At the end of the trial, 22 volunteer curators responded to an open-ended questionnaire regarding their experiences with the platform. Further semi-structured interviews were held with five of these curators. Interview questions followed up on the free-text answers from the questionnaire and focused on usability, workflows, motivation, perceived value of the work, and platform extensions.

Seven domain experts with extensive experience from the disaster management domain were also recruited to explore the curated stories in CrisisTracker during the field trial week and assess the value of the system. Their expertise included disaster management at different levels, GIS analysis, information management and media monitoring. Two of the experts were actively following the Syrian conflict beyond the scope of our study and three experts also participated as volunteer curators. Semi-structured interviews were held with all seven experts with questions relating to 1) their past use of social media reports during crisis, 2) their knowledge of the Syrian conflict, 3) their experiences using CrisisTracker, 4) potential applications of CrisisTracker during past events, and 5) its ability to support different users and use cases.

Transcripts from interviews, plus chat logs, survey answers, emails and other communication were content analyzed by the author through bottom-up coding, clustering and interpretation (Hsieh and Shannon 2005).

To collect tweets, a bounding box was defined covering Syria and together with an analyst familiar with the conflict, 50 keywords were selected to track on Twitter in English and Arabic (syria, assad, hama, damascus, aleppo, homs, bashar, #fsa, daraa, #siria, #syriarevo, #syrie, #syrian, #syrias, idlib, #realsyria, lattakia, houla, latakia, rastan, deraa, #syr, hamah, tartus, harasta, zabadani, darayya, baniyas, babaamr, darayaa, raqqah, yayladagi, basharcrimes, suweida, latakiyah, دمشق , حمص , #الجيشالحر , #جيشتويتر , #حلب , #حماة , #الجيشالسوريالحر , #بشار , #معركةدمشقالكبرى). The words referred mainly to major place names and popular hashtags. The same analyst also assisted in defining 14 relevant report categories to be applied to stories by the human curators, e.g. “demonstration”, “eyewitness report” and “people movement”. About 3.5 million tweets were collected during the evaluation week, but the system had been up and running for several months before the volunteers began working and older stories could be retrieved through searches.

Only a few participants spoke Arabic and curators were encouraged to use machine-translation features built-in to some Web browsers to read stories and linked articles, and add meta-data based on the translated content. The platform supports inclusion of manually translated summaries to stories, which native language speakers in some cases provided. A user interface feature was also implemented to hide Arabic stories for users who found the Arabic content too tiring or difficult to work with.

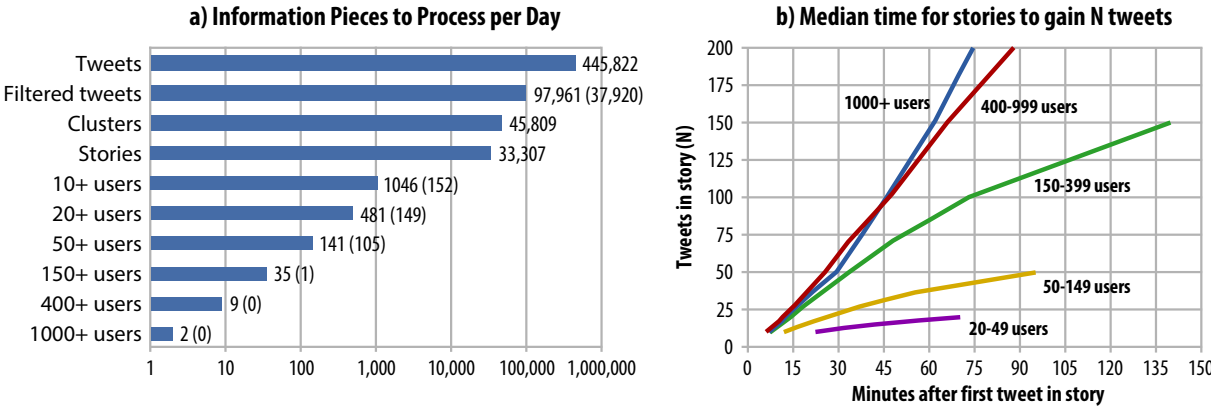

FIGURE 6.4: (a) Reduction of information inflow rate from each processing step. "N+ users" refers to stories containing tweets from at least N unique users. Numbers in parentheses indicate additional items produced by the 58 spammer accounts. (b) Growth rate of stories. If stories are filtered by size, impactful events can realistically be detected around 30 minutes after the first tweet.

6.4 Results

6.4.1 Clustering reduces workload

Clustering the incoming tweets into unique stories greatly reduced the rate at which new items needed to be processed. The system collected on average 446000 tweets per day (min 400783, max 560193). 70% of tweets were immediately discarded, almost exclusively because the geographic bounding box covering Syria overlapped with a small part of Turkey, causing geotagged tweets from anywhere within Turkey to be returned by the API. The remaining messages were then clustered into about 33000 daily stories, of which 1200 stories contained tweets from at least five users and 246 stories from at least 50 users.

Although size-based story ranking has been shown to generally improve signal-to-noise ratio and to remove some forms of spam (Starbird et al. 2010; Rogstadius et al. 2011a), high levels of spam were initially found among the top stories. These spam stories mainly originated from a group of 58 highly active spammer accounts. Once these accounts were blocked, the number of daily stories shared by at least 50 users dropped to 141. Focusing on these top stories thus reduced the workload by three orders of magnitude, with only a handful of new top stories per hour. Figure 6.4a summarizes how each processing step contributed to making the workload manageable.

6.4.2 Scalability of crowd curation

The fully automated components of CrisisTracker collect and cluster tweets into ranked stories, and extracts sufficient meta-data to support search and filtering based on time and keywords.

Human curation is required for location-, category- and named entity-based filtering, which helps find reports that get less attention by the Twitter user base.

Curators spent on average 4.5 minutes per story, with a heavy skew towards shorter times (median 2.3 minutes). Each work session lasted on average 28.5 minutes (median 20.8), with work sessions defined as periods of user activity separated by at least 15 minutes of idle time. A total of 3600 tags were added to 820 stories (1775 before merging), and together the curated stories contained 616009 tweets.

This total volunteer output can be considered substantial when compared to most Ushahidi deployments. Data provided by CrowdGlobe[1] shows that out of 871 cloud-hosted Ushahidi instances[2] that were ever active (containing ten or more reports), only 67 (7.7%) contained over 820 reports. Furthermore, 14 of the 15 public instances with 500 to 1100 reports received their reports over more than two months, compared to the eight-day effort in this study. As curation tasks are similar in the two platforms, we mainly attribute the higher productivity of CrisisTracker curators to automated information gathering.

In agreement with the findings in chapter 4, the work effort was unevenly distributed among the participants and 25% of the curators did 75% of the work. Among the 22 people who curated at least 10 stories each, the average time spent per person per story ranged between 1 and 13 minutes. Only in a handful of cases did curators work on the same stories or remove others' tags.

Based on the reported usage statistics and the rates at which information was collected, we estimate that about 15 curators each active for 30 minutes per day would be sufficient to have full meta-data for all the main events during a humanitarian crisis of this type and magnitude. Approximately 150 curators would be able to build and maintain a very detailed database of almost all reported events, below city-level resolution. Workforce size can be greatly reduced if curators can be encouraged to spend more than 30 minutes per day. Feedback from volunteers also suggests that per-curator effort would have been higher if the deployment had been a request from a humanitarian organization, rather than an academic study.

6.4.3 Real-time overview

One of the domain expert participants, an anonymous English and Arabic speaking representative of Syria Tracker[3], was actively monitoring ongoing events in Syria in parallel with using CrisisTracker. Syria Tracker had been monitoring the crisis for 18 months and had set up infrastructure to automatically mine Arabic news media. They also received daily eyewitness reports

[1] http://crowdglobe.net/
[2] https://crowdmap.com/
[3] https://syriatracker.crowdmap.com/

via email from trusted sources on the ground, and manually monitored online social media. This participant was therefore in a unique position where he could in real time compare the information made available through CrisisTracker with that independently collected through other methods.

Known in advance was that Twitter was from the start of this conflict an active communication channel used mainly by the opposition, and more recently also by government supporters. Links to relevant images, videos and news articles were also often posted on Twitter almost immediately following publication of the resource. Syria Tracker was well aware of this rich source, but had not found any way to reliably monitor it in real-time. After using the system, the expert explained that CrisisTracker was the first of many tools they had tried that provided a subjective sense of real-time overview of this "information storm", in both English and Arabic.

A GIS expert who was also one of the most active curators described how seeing real-time commentary on what was happening in Moscow, next to reports of events on the ground and how many people were killed gave her a sense of how the whole world was connected. She further described how "you can see over a period of time where people are moving, how that relates to conflict areas. Water shortage, or food, you can almost anticipate where needs are going to be." As discussed later, reaching this level of understanding requires a significant time investment. However, we consider it highly promising that such levels of situational awareness can at all be gained from data largely derived from automated processing of freely available social media.

6.4.4 Timely and rich reports

According to Syria Tracker, the tool improved sensitivity of information collection by helping them detect several important events (e.g. massacres, explosions and gunfire) before they were reported by other sources. CrisisTracker also improved specificity, by quickly picking up links to evidence such as images and videos. Early detection of events enabled focused manual monitoring of other sources to corroborate evidence and make early assessments of the truthfulness or severity of new claims. For instance, videos of missiles being fired with claims regarding time and place led to searches for corresponding impacts.

One of the concrete ways in which Syria Tracker used the system was to take trusted but brief and single-sided eyewitness reports submitted via email, and in CrisisTracker quickly seek out complementary pictures and videos as well as reports from the opposing side. This augmentation of eyewitness reports enriched their understanding of ongoing events hours and sometimes even days before mainstream media picked up the news. During ongoing events such as urban skirmishes, CrisisTracker would also provide real-time updates whereas eye-witness reports

would arrive as high-quality summaries later during the day. Syria Tracker also valued the historical record of social media communication provided by the tool, as Twitter does not support searches more than a few days back in time.

Six of the seven interviewed experts in some way mentioned timeliness as a primary reason for using the system. An incident commander who actively used the platform over several days said *"I feel very confidently that those reports will come out ahead of CNN and BBC and that they will have the central nuggets of who, what, when, where, why. For an incident commander, it is the difference between learning something in two to three hours versus learning it in six to eight."* A data analyst with a background in disaster response further added that the timeliness and being able to see how stories are evolving is *"huge."*

Figure 6.4b shows how quickly stories of different magnitude would be detected when monitoring all stories above different size thresholds. This graph is informative in the context provided by Figure 6.4a, as there is an inherent trade-off between how quickly the system can detect important news and how many such stories the system will detect per day. If set too low, this filtering threshold will simply cause information overload. A size threshold set at 50 users would in this study have produced around 141 stories per day, which can be considered relatively manageable. Among a sample of 9207 stories that eventually contained tweets from 150 or more users, and which can be considered to capture significantly impactful events, 10% would be detected within four minutes, 50% within 31 minutes, and 90% within 3.5 hours of the first report. It is our impression that the rapidly detected stories to a greater extent cover instantaneous events such as explosions, whereas the more slowly detected events refer to less critical news, such as international politics, but further research would be needed to verify this.

On its own, Twitter is in fact so timely that several participants expressed past feelings of frustration regarding how difficult it is to keep up with the media. One participant stated, regarding tools he has used in the past, that *"if you're not on the Twitter stream you quickly lose sight of the history. It becomes very much a snapshot, a moment-in-time assessment of what's happening"*. Since CrisisTracker keeps a historical archive of all larger stories as the main event progresses, it becomes possible to go back in time and analyze both short-term and long-term trends.

6.5 Discussion

6.5.1 Usage barriers

Arabic content proved frustrating for many non-Arabic speakers, who simply hid it and focused on stories in English. Others were concerned about the quality of machine translations, but it

was noted that "obviously some reports seem just fine with little room for error, while others are just unusable." Another volunteer said that *"using Google Translate was very easy and by being able to curate stories, especially those in Arabic, I felt more of a 'direct' connection with what was going on."* One of the greatest obstacles interestingly proved to be geotagging of English-language reports, as the transliterations of location names often could not be found on the map or through search. Overall, our impression is that machine translation is tiring but relatively safe to use for curators, as the impact of mistranslation is limited to tagging errors that reduce the quality of search and filtering.

Many experts and volunteers made remarks about the system's overall intuitiveness and ease of use. A few users explicitly compared CrisisTracker with their experience using the Ushahidi system, which they considered to be more tedious to work with and less intuitive. Participants with experience from both systems specifically valued CrisisTracker's ability to relate stories to each other and its method of handling duplicate reports. However, the evaluation also identified several minor user interface issues that lead to breakdowns in user interaction and which should be resolved before the platform can be considered mature enough for public use.

In rapid onset disasters, ease of deployment (effectively deployment time) is a critical aspect of an information management system. While anyone can download the CrisisTracker source code and set up an instance, further work is needed to simplify and automate the deployment process. Ideally this would be done through a system similar to the Crowdmap website[4], where new instances of the Ushahidi platform can be deployed on cloud servers through a web interface.

Many participants raised concerns that the system's complete openness combined with lack of "undo" functionality leads to high sensitivity to vandalism, in the form of purposeful incorrect tagging, merging or hiding of stories. This is particularly an issue during conflict settings and extensions need to be made to handle this, either by adding mechanisms for screening of curators, or by implementing a version control and community moderation system similar to those on Open Street Map[5] and Wikipedia[6].

6.5.2 Using CrisisTracker to support decision making

This evaluation suggests that the greatest value of CrisisTracker during complex and constantly changing large-scale events is improved real-time situation awareness, at the perception, comprehension and to some extent projection levels. This result is very similar to that of an evaluation study of the Ushahidi deployment for the 2010 Haiti earthquake (Morrow et al. 2011), which showed that improved situation awareness at an aggregate level was the greatest contribution

[4] https://crowdmap.com/
[5] http://www.openstreetmap.org/
[6] https://www.wikipedia.org/

of the crowdsourced map, in particular during the early days of the crisis when the situation on the ground was still unclear. Content analysis of tweets collected during natural disasters (Vieweg 2012) indicates great availability of response-phase related reports: hazards, interventions, fatalities, personal status and damage. This agrees with our general perception of the content distribution for the Syrian conflict. Though one disaster manager participant in our study speculated that CrisisTracker would be of great value also in the recovery phase, the content analysis indicates that recovery-related tweets are scarce.

CrisisTracker also addresses two important limitations identified by the Ushahidi evaluation. First, CrisisTracker taps into existing social media to access the voices of affected populations, rather than relying on independently submitted or manually collected reports. Second, while the vast majority of reports go uncurated in both CrisisTracker and Ushahidi, CrisisTracker is able to direct curation efforts towards those reports that are discussed most in the core community and thus most likely to improve situation awareness.

Despite general praise for CrisisTracker's good usability, intuitiveness and ability to extract rich and timely information of important events, many participants made clear that gaining an accurate understanding of the information remains a time-consuming task. A high-level manager in a humanitarian organization explained that while much useful information is collected by the system, key points need to be distilled from the stories to make that information useable in time-pressed situations.

The system is therefore not yet ready to be used directly as a decision support tool by decision makers who have very limited time to sit down, read and analyze information. Rather, CrisisTracker is suitable for use by analysts and others who already work on filtering and aggregating information from different sources to produce maps and reports tailored for the organization's decision makers. The interview subjects who worked in analyst roles were very enthusiastic about the tool and only requested export functionality to be implemented, to enable comparison of social media reports with data from other sources.

Several participants, both curators and experts, though none of those with analyst backgrounds, expressed that they wanted CrisisTracker to help them assess the trustworthiness of stories in the platform. Discussions focused on that assessment should be done on a source level, for instance by inferring how credible a first report is based on the past record of that account, or by highlighting the most trustworthy account that has shared a story. Others noted that different sources have different authority for different information, which greatly adds complexity to such calculations. For instance, a personal account with strong political bias may contribute little to international news, but may still be authoritative regarding events in the particular suburb where they live. In other cases an untrusted source may post a link to a highly trusted government website, or a trusted government account may post outdated information that is contradicted by citizen generated video footage. Potentially feasible approaches include letting

curators mark single accounts as having extremely high or low credibility, and deriving source ratings from sharing patterns of past stories.

One participant noted that human curation itself may be misinterpreted as validation, which if true would be a serious risk in decision making and could be an entry barrier for volunteer curators. This is something that future development will need to keep in mind and find ways to avoid.

Several of the more experienced disaster analysts and managers we interviewed, or have spoken to at other times, have pointed out how accuracy and timeliness will always remain a trade-off. Verification requires both time and expertise, and a system that can deliver very quick reports in a consumable format does not need to always be correct to be valuable. One disaster management consultant estimated that even her trusted personal sources that she uses for verification may only be right in 80% of the cases. Several participants also reported that they felt the system's grouping of complementary reports into stories helped build reliability and supported validation. Based primarily on the contextual feedback provided by Syria Tracker, we strongly believe that the rich and timely but unverified reports in CrisisTracker are most valuable when combined with other sources, for instance to enrich brief but trusted eyewitness reports. We also believe the system is capable of providing rich historic narratives around specific points in time and space that can help explain interesting features in other datasets.

6.5.3 Managing a CrisisTracker deployment

When integrating crowdsourced human-based computation into disaster information management, leadership has a direct impact on performance metrics such as precision, recall and processing speed. Based on experiences from chapter 4 and from this study, we propose that humanitarian organizations that want to deploy this class of systems assign a person the new role of crowd director.

First, the crowd director is responsible for recruiting curators, e.g. the organization's own registered disaster volunteers, or through an independent volunteer organization such as the Standby Task Force. As volunteering competes with other tasks in people's lives, it is of great importance during recruitment to motivate potential curators by transparently explaining how the work will create benefit and help a population in need. Volunteers will also get up to speed quicker if they receive basic background information regarding the disaster. Basic training materials regarding how to use CrisisTracker are already in place.

Once work begins, the crowd director must continuously communicate decision makers' information requirements to the crowd. This will enable the organization to steer the crowd towards work that is of particularly high value. For instance, instructions can be to focus on reports

relating to a particular town or report category, or to spend more or less time on tracking down precise locations. Unlike information management systems that are completely automated, interactive crowd management enables CrisisTracker to effectively function differently depending on the particular needs of each deployment. Rather than having to fine tune or develop new processing algorithms for each new use case and content language, organizations can themselves maximize performance by giving direction and recruiting volunteers with relevant skills and experience.

Finally, insecure or inexperienced volunteers will ask for affirmation that their work is correctly carried out. The crowd director must remain accessible to provide such feedback, which can be the difference between having a volunteer who only curates a single story and then stops, and one who confidently comes back day after day during most of their spare time. Some intervention may also be required to improve accuracy of the assigned labels, by prompting volunteers who despite their good intentions make mistakes or too rushed classifications (e.g. placing geotags in the middle of cities when more specific information is available).

6.6 Conclusion

The goal of this study was to understand what value a tool like CrisisTracker can provide during a humanitarian disaster. Based on a case study in a conflict setting, we see strong indications that a hybrid analysis process can be an economical way to greatly improve the real-time situational awareness of an analyst team with limited resources. While the tool complements rather than replace other methods of information collection, its timely and reliable event detection helped the participants direct their attention to where it was most needed.

Based on expert feedback collected during the study, a social media monitoring tool like CrisisTracker is best targeted at analyst roles, as one information source among many. During its use, CrisisTracker picked up both fine-grained single events and coarse high-level summaries, which all contribute to greater comprehension of how events at micro and macro level fit together. However, the system currently has no way of reliably extracting quantifiable measurements of humanitarian impact, nor can it provide any form of automated credibility assessments for summaries. While the current feature set is sufficient to be useful for analysts, it will likely remain difficult in the foreseeable future to design tools to be directly used by time-pressed decision makers in the humanitarian domain.

Naturally, the utility of the tool is entirely dependent on the relevance and volume of citizen communication in social media. While previous research has indicated that such reports are prevalent during the response phase of disasters, there are no clear indications that the same holds true for the mitigation, preparedness and recovery phases. As the system only gathers

public communication, it is also unlikely that it will ever be good for collecting reports about stigmatized humanitarian suffering, such as rape. Further content analysis of social media communication is needed to investigate both of these aspects.

The utility of the current system is also greatly affected by the status of communication infrastructure, which may be destroyed during a disaster. However, future versions may be extended to globally monitor geographical pre-disaster communication patterns to detect the *absence* of reports and use this signal as an early indication of severe infrastructure damage. It is also likely that the resilience of communications infrastructure will increase in years to come, though to what extent is difficult to say.

A limitation of the clustering algorithm currently used in CrisisTracker is that it comes with a significant start-up time, during which the system is learning the relative frequencies of words used during a disaster. The resulting processing delay, which affects clustering but not report collection itself, may limit the system's utility during the early hours or even days of rapid onset disasters. Future work can look at how to improve this aspect, potentially by using a different clustering algorithm, or by bundling the system with a default dictionary.

Finally, while CrisisTracker's automated de-duplication of reports and ranking of stories improved the effectiveness of crowdsourced human-based meta-data extraction, there is no way that human workers can keep up with and annotate *all* the valuable information in the stream through direct annotation of each story. As high-quality and high-coverage meta-data is a prerequisite for many extensions to the system, in particular visualization interfaces, the next chapter will discuss ways in which the system can learn from the tagging behavior of human curators, to provide a best-guess for content that no human has processed. The next chapter will also indirectly address the current system's sensitivity to vandalism and unintentional curation mistakes.

Chapter 7

Towards scalable meta-data extraction

The CrisisTracker system, presented in the previous chapter, combined automated and human computation into a hybrid system capable of summarizing and structuring citizen reports from social media during an ongoing disaster. By clustering social media messages into unique stories, the system brought overview and helped detect new events early and reliably, contributing to general situational awareness.

However, highly specific content filters are required to support decision makers with narrow areas of interest, such as logistics, food or shelters. Furthermore, structured information needs to be extracted in high volumes and with great accuracy before it is possible to perform any meaningful quantitative analysis to detect high-level trends and patterns in large sets of information. In the previous chapter this was provided through meta-data, extracted from stories through human computation. Meta-data based information filtering is akin to the proverbial needle in the haystack problem; finding information of a specific type in real-time within a rapidly growing stack of information. To identify valuable information, and to direct information to the right decision maker, messages need to be sorted into high-level, meaningful categories of information.

Through its automated de-duplication and ranking of reports, CrisisTracker improved efficiency and effectiveness of human computation over previous comparable systems used in the disaster information management domain. However, the system architecture did not address the fundamental issue of increasing the total work output of a given crowd. In both Ushahidi and CrisisTracker, an average human annotator will process at most a few dozen items per hour. While this may be sufficient to annotate the top stories capturing highly impactful events, the hourly inflow of crisis information posted on Twitter during major disasters can total tens of thousands of tweets or thousands of new stories. Direct human annotation can thus never be scaled to process more than a fraction of the total content. Personal experience from extensive use of CrisisTracker has concluded that hidden in the long tail of unprocessed single reports is

a vast body of messages that may not contribute much to the understanding of the disaster as a whole, but which would be of use in supporting specific decisions if they could be matched with equally specific search criteria.

Even if a large enough workforce could somehow be amassed to keep up with the high velocity of crisis information, having humans do something that can be automated is inefficient and a misallocation of limited resources. This is especially true during prolonged crises, such as civil wars, or during the long recovery phase following major natural disasters. In these cases, dependency on continuous high-volume human annotation adds a significant operating cost that cannot reasonably be expected to be sustained over time.

One approach to achieve scalable meta-data extraction would be to replace human curators with tailor-made and pre-trained machine classifiers. However, automated classification of microblog content is significantly harder than for longer documents such as news articles or blog posts, due to, among other things, great language variance and feature sparsity (resulting from limited message length). In addition, Imran et al. (2013b) found that classifiers trained on social media content collected during one disaster suffered greatly reduced classification performance when applied to data collected during other disasters, even when the disasters were of the same type and affecting the same country. Because of these issues, it remains difficult to build systems that incorporate pre-trained classifiers to be deployed during new events.

7.1 Supervised learning of stream classifiers

Imran et al. (2013b) however found that automated classifiers trained on human-annotated data, sampled from a corpus collected during a specific disaster, performed reasonably well at classifying other reports from the same corpus. This finding suggests that it may be possible to apply supervised learning techniques during an ongoing human-based classification effort to generalize the tagging behaviour and increase the processing capacity of the human crowd.

This chapter describes such a system, built for real-time meta-data annotation of social media streams, through iterative supervised learning of stream classifiers. Using this approach, end users can develop automated classifiers tailored to the information needs, language and characteristics of reports in a specific disaster, rather than being dependent on or limited to pre-defined generic classifiers. This first version of the system supports assignment of one or multiple labels from a finite set of user-predefined labels, which can represent for instance different humanitarian needs sectors that apply in the disaster of interest.

Because of the high rate at which information is generated, the classification system must be able to sustain a high rate of throughput. To achieve this, the classification system consists of multiple parallel processing steps, connected into a stream processing pipeline (Figure 7.1).

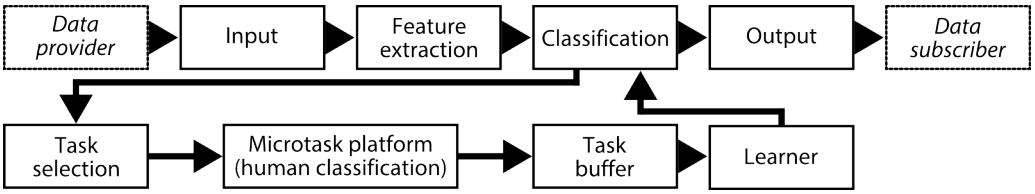

FIGURE 7.1: Flow diagram of the classification module's architecture.

The system described here was developed as one module of the AIDR service, developed at Qatar Computing Research Institute[1]. AIDR's architecture, including the classification module, has been described in detail in (Imran, Lykourentzou, and Castillo 2013), in which interested readers also can find a quantitative assessment of the system's good throughput capacity, load adaptability, cost effectiveness and output quality.

Each pipeline step is connected with the next by a queue data structure, to make the system more robust to spikes in the message input rate. If at any point new messages arrive at a higher rate than they can be processed and the size of any queue grows past a threshold length, new messages are discarded instead of being appended to the queue.

7.1.1 Classification

To begin processing, a data provider, such as the module in CrisisTracker which consumes the Twitter stream, establishes a persistent connection to the input step of the classification system. This input stage parses received messages into an internal data structure, then pushes the parsed message onto a queue and waits for the next message.

Next is feature extraction, which takes a message from the queue, performs stop word removal of the message text, extracts word unigrams (single words) and bigrams (word pairs), appends these to the message, and pushes the item to the next queue. Additional feature types can be added in future versions of the system, but previous work (Imran et al. 2013a) used word unigrams and bigrams to reach acceptable classification accuracy for the tested classes.

After feature extraction, the message is passed on for classification. Here, a set of currently active classifiers (more on this below) is kept in memory, and each classifier is applied to the message to generate an output label with an associated confidence score. These are appended to the message, which is then passed on to the output step, to which consumer applications can subscribe to receive a stream of classified messages.

The system supports classification of items into sets of categories, where each set consists of one or many labels. In addition, each set also contains a null label, which represents that the classified item is not similar to any training data available for any of the labels in the set. If no

[1] http://aidr.qcri.org/

human-labelled training data has been provided yet, no classifier will be available and all items will be assigned a null label with minimum confidence.

7.1.2 Training

After classification, each message is also passed to a task selector, which is responsible for identifying messages that, if labelled by a human, are likely to provide new knowledge not currently represented in the currently available classifiers. First, it looks at the confidence scores of the labels assigned to the message and discards it if the confidence scores are high, meaning that the message has highly similar features to those of the messages already in the training set used to produce the classifiers. This process is commonly referred to as active learning. The task selector then performs rudimentary de-duplication using simple heuristics (a time-sorted buffer and word set overlap comparisons) to remove messages that are highly similar to other recently processed items. As the classification system is currently applied to messages from Twitter, re-tweets are discarded in this step. In the future, the de-duplication step can be further integrated with CrisisTracker to make use of its more advanced clustering algorithm.

If the message is sufficiently novel according to both metrics, it is passed on to a task buffer. This buffer is connected to a basic custom microtasking platform, in which human annotators can assign high-quality labels to messages that the system does not yet know how to classify. Each human-annotated message is then pushed into a database of human-labelled messages, either directly or after receiving a sufficient number of agreeing votes from different users. In the database of human-labelled samples, 20% of messages are put aside for model evaluation, while the remaining 80% are used for model training.

As new messages are added to this database and the available training set grows, the system automatically applies a supervised learning algorithm (currently a random forest as implemented in Weka 3.7.6[2] to train new classifiers. As new classifiers become available, they replace the previous active model for that message class in the classification pipeline.

An administrator of the system can at any point in time inspect the estimated classification accuracy of the active model through a web interface, as well as see how performance has changed with different numbers of available training samples.

[2] http://www.cs.waikato.ac.nz/ml/weka/

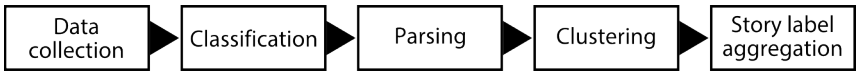

FIGURE 7.2: Flow diagram of the classificaiton module's integration with CrisisTracker.

7.2 Integrating supervised classification in CrisisTracker

7.2.1 Story classification

The classification system was integrated as a processing module in CrisisTracker. Primarily, a decision had to be made if classification should be done before or after clustering, as the end goal was to assign topic labels to stories rather than individual messages. Performing clustering before classification would make it possible to classify stories based on a much richer set of information than what a single message can provide, as the field study showed that related messages in a cluster tend to capture different complementary aspects of an event. Classifying entire clusters would also reduce the number of items that need to be classified. However, this approach is difficult to take in a streaming environment. When real-world events continuously change and unfold, cluster contents keep changing and clusters would need to continuously be re-classified.

Instead, the integration was done such that all messages in the stream are first classified independently, and the labels are stored in the database along with the other message data. This process flow is shown in Figure 7.2. Null labels output by the classifiers as well as labels with confidence scores below a threshold value are discarded and these tweets are stored without labels. Next, messages are clustered based on their textual content, in the same manner as in the previous version of the system.

After each second-order clustering pass, in which tweet clusters are grouped into stories, all stories that received new tweets since the last pass are re-classified by finding the most common (pre-calculated) label within each of the label sets, among all messages in the cluster. This majority voting is a cheap operation that can be done through a simple database query, as opposed to a full re-classification of the story based on its updated content.

The end result is that a story matches a label search if the most common label is the search label, when only considering those tweets that the classification module could assign a non-null label with a high confidence score. Within label sets, search criteria are combined using logical OR, and across label sets, search criteria are combined using logical AND.

7.2.2 User interface

Along with the addition of the classification module, CrisisTracker's web front-end was redesigned to match the new way of handling meta-data, as well as to streamline the user interface around the system's core strengths. As identified in chapter 6, the system's greatest value to end users came from the timely and rich summaries of emerging and ongoing events that are provided through its fully automated clustering algorithm. While meta-data annotations helped analysts filter the content, direct human annotation is very likely to be too costly to be reliable in prolonged humanitarian disasters, such as civil wars. Therefore, the previous filtering dimensions (time, geographic locations, topics, named entities and keywords) was reduced to filtering by time, classifier-provided labels and keywords.

Previously, CrisisTracker provided one page for navigating between stories, one page for exploring the content of a story and adding meta-data annotations, and one page to assist curators in selecting what stories to work on. These pages were all replaced with a single page for navigating both stories and the raw tweets contained within. Training data for the classification module is provided through a separate page, solely focused on labelling system-selected tweets with labels selected from administrator-defined label sets.

The new front-end, show in Figure 7.3, consists of three vertical panels, labelled Filters, Stories overview, and Focus story. An end user can use the Filters panel to retrieve stories by selecting one or several topic filters, or entering one or multiple search keywords. The search can also be limited to a specific time interval using the timeline at the top of the Stories overview panel. The search then returns a list of matching stories in the centre panel, showing the time of the first tweet within each story, the number of tweets within each story, and each story's title. Results are sorted in reverse chronological order. If a time filter is applied, the search returns stories first seen within the time range, ranked by their number tweets from the top-5000 users, otherwise the search returns the stories that received the most new tweets from the top-5000 users in the past four hours.

Selecting a story in the centre panel brings up its content in the right-most Focus story panel, where the user can also narrow down the search further using time, keyword and label filters to find specific messages. Within a story, messages are ranked by a weighted product of the number of near duplicates and the distance to the nearest neighbour, as determined by the clustering algorithm, with higher ranks given to novel messages with many subsequent near duplicates. This provides a summary that gives priority to the most popularly shared version of the story, while capturing as much diversity as possible. Similar stories are listed in a separate tab in the Focus story panel, from which a user can navigate to and merge related stories.

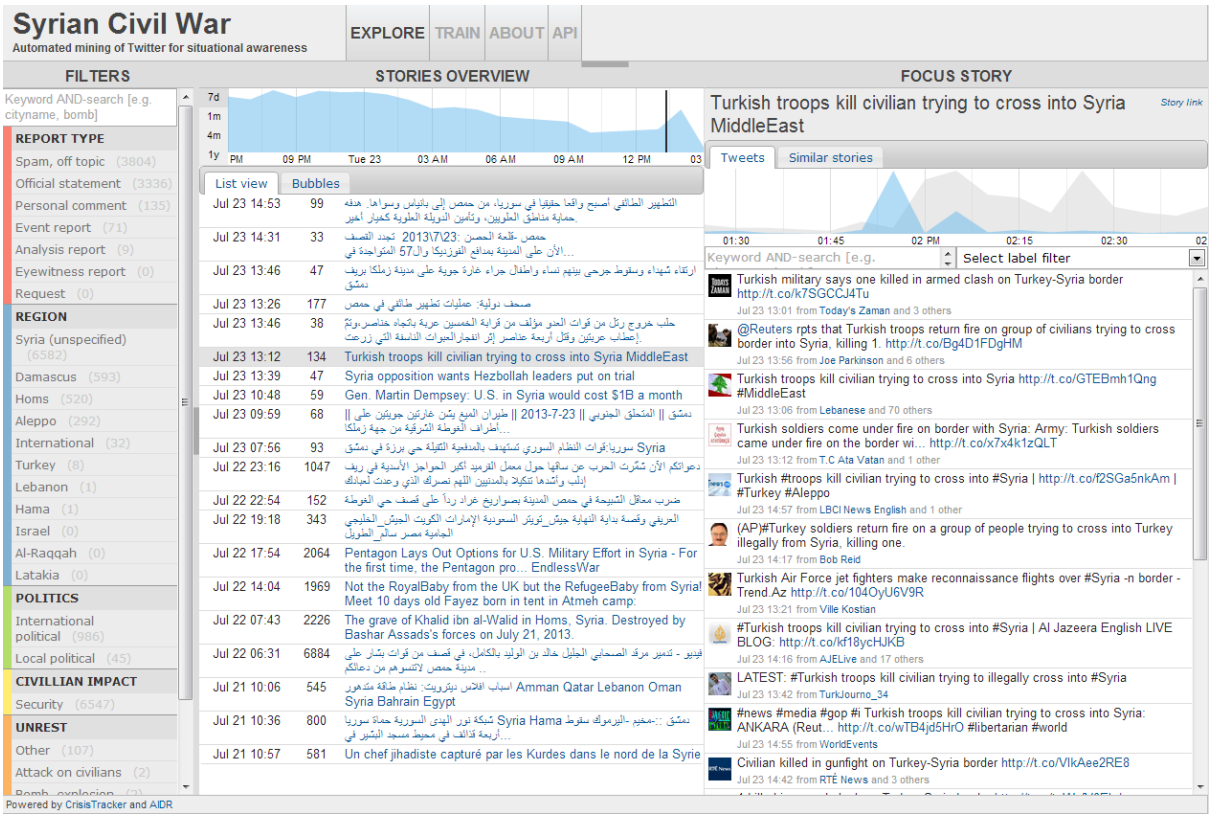

FIGURE 7.3: The updated CrisisTracker web front-end, which includes customizable topic filters (left) and time filters (top centre and top right). Topic classifications are provided by the new integrated classification module.

Selecting an individual message from the list in the left-most panel brings up a full rendering of the tweet, including an embedded preview of link targets, such as a video, image or news article. This expanded tweet view is provided by Twitter through its API.

7.3 Limitations and remaining challenges

The classification system as described in this chapter is a work in progress, and in addition to the evaluation of AIDR conducted by QCRI, the classification module has been tested informally by the thesis author in a CrisisTracker deployment focused on the Syrian civil war. This evaluation has subjectively looked at how the classifiers perform after integration with CrisisTracker and has so far uncovered a number of limitations in the classification system. These experiences are documented here along with reasoning regarding their underlying causes, with the hope that these lessons learned may help inform future design decisions in projects aiming to develop related systems for online supervised learning.

7.3.1 Classification of infrequent or anticipated labels

The active learning process in the system uses classification confidence to prioritize which of the many incoming documents that are selected as tasks for a human to label. This is an important process, as the number of possible documents to select is very large, and the quality of future classifications will be based entirely on the representativeness and coverage of the hand-labelled training data in terms of news topics, textual features and assigned labels.

The active learning process seems to work fairly well to refine a decision boundary to distinguish between classes, as well as to achieve approximately even sampling for each output class, but it has a significant limitation. It can only be applied once there are a sufficient number of examples of each output class for the system to be able to train the first model.

Before any training data is available, and thus no classification model has been trained, the system has no good way of knowing which messages are likely to be more informative in model training, and task selection is essentially random. When class distributions are balanced, random sampling quickly produces examples of each output class. However, random sampling is very inefficient when class distributions are skewed, either in absolute frequency or in time, as can be the case with classes that have not yet been seen but which are anticipated to appear in the near future.

If $p_L \in [0, 1]$ is the probability of randomly selecting an item with desired label L, and N is the number of positive examples that need to be labelled before a classifier will have non-zero classification accuracy for that output class, then a human needs to label approximately $M = N/p_L$ randomly selected documents before the active learning kicks in. Based on informal evaluation, N is around 10 in the system, thus if $p_L < 1\%$, which is not uncommon, then more than a thousand documents may need to be hand-labelled before any model is trained and active learning is activated.

It is important to note that this issue is independent of classification performance. Even after labelling thousands of examples, the first model trained will only have a handful of positive examples to generalize from, and even more labelling effort is needed to train the classifier to acceptable output performance. There are several ways in which this issue could be mitigated.

First, features could be added to completely eliminate the need for pre-model task selection, for instance by letting an administrator provide synthetic examples for the labels they have defined. A synthetic example is an artificial (or copied) message, which is representative of what future messages of interest will look like. While writing such examples is quite challenging and likely time consuming, the required workload would be far less than labelling a random selection of documents for low-frequency classes. In addition, synthetic examples have the advantage that they can be prepared in advance of a slow onset or anticipated disaster, such as a hurricane or

earthquake, so that basic classifiers are in place already when the first reports start coming in. Synthetic examples could also be requested from crowdsourcing workers once it is clear that a label is rare.

Alternatively, one could alter the pre-model stage by replacing random sampling with some form of guided sampling. The system could let an administrator specify several keywords that they believe will correlate positively with each sought label. These keywords would then be used for stratified sampling during the initial phase, to sample up to some number or ratio of messages that are somewhat likely to belong to each class. The benefit of this approach is that it takes less effort to write down a number of related keywords, than to provide a varied range of synthetic examples that are representative of real communications. The keywords can also more easily be reused for different crises, as they are less contextual than synthetic samples. In addition, the training corpus will only be made up of real documents, which minimizes the negative impact on classification performance from accidentally providing synthetic samples that contain features that are unrelated or ambiguous.

Finally, it would be possible to maintain a centralized repository of human-labelled training sets, collected in different languages during past disasters of different types. As the repository grows, it would increasingly become possible to seed new deployments with pre-labelled data from past similar disasters. This approach however imposes a strong limitation, namely that label definitions (the criteria that human annotators rely on to assign labels) remain stable between disasters. The approach may thus be attractive only for information categories that remain highly consistent between disasters. If label definitions can be reused, the effect of keeping a repository of past data would in fact offer similar strengths as providing pre-trained classifiers, but with the additional significant benefit of being able to further improve classification performance beyond off-the-shelf levels during an ongoing event.

7.3.2 Maximising real-time classification performance during peak activity

The overarching purpose of the work presented in this chapter has been to provide real-time classification of citizen communication during ongoing humanitarian disasters. Slow and rapid onset disasters each have their associated characteristic pattern of communication. Rapid onset disasters such as tornados or earthquakes consist of a short destructive period, followed by a much longer reconstructive period. Total volume of communication will match this pattern, with a large spike immediately following the initial destruction, while message volume then gradually declines during the response and recovery period. This is illustrated for instance in Figure 1 in (Imran et al. 2014), in which spikes in message volume last for one to a few days.

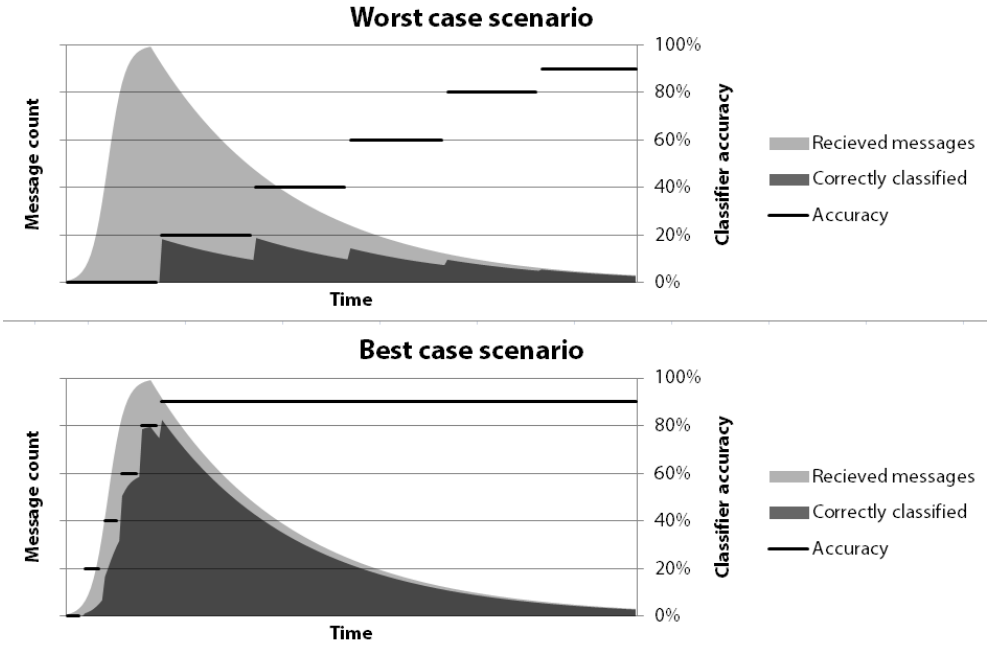

FIGURE 7.4: When performing classification in real-time during rapid onset disasters, it is crucial that classifiers are trained to sufficient accuracy before the bulk of the communication takes place. If not, only a small fraction of the communication will be correctly classified. In conflicts and other long-lasting crises, message inflow rates follow a different pattern and classifiers can be trained over longer time periods. These graphs are illustrative and do not represent actual data.

This raises two questions. First, when in time are the messages belonging to a specific class of interest posted? Second, how accurately will system be able to classify reports of that class at the time they are posted?

With the current system implementation, classification accuracy for each class will be zero at the time when the first message belonging to that class appears. As more messages are being posted, human annotators will provide training data, and gradually new classifiers will be trained with increasing accuracy. How quickly this learning process takes place in real time depends both on the number of human annotators that work at the given point in time, and on how representative the early messages are of future content. Figure 7.4 illustrates both best-case and worst-case scenarios in which sufficient training data for accurate classification becomes available either before or after the spike in communication during disaster onset.

The best performing classifier for past data at a given time will be that which has been trained with all available training data, but since training data is collected incrementally, this classifier will currently only be applied to future data. All data collected up to this point will have been automatically classified using some worse model. It would in theory be possible to gradually re-submit past data for re-classification, but this would result in a situation where the workload continues to grow with time. System performance would thus eventually start to degrade, unless the system could somehow determine which subset of past items to re-classify. To complicate

matters further, if reports are clustered and clusters are indirectly assigned aggregate labels based on the labels associated with the items within the clusters (as in CrisisTracker), the entire cluster's content may need to be re-classified before the aggregate value will change. Use of historic training data from past disasters can ensure that classifiers are available from the disaster onset, but as previous research has shown, past training data alone cannot be assumed to produce highly accurate classifiers.

Additional research is needed to investigate how these issues can be mitigated, to improve the utility of supervised learning systems in rapid onset disasters. The system in its current state likely has greatest value in slow onset natural disasters or man-made hazards such as civil wars. These events can last for long periods of time, and having access to high-performing classifiers from the very first day is less critical.

7.3.3 Handling accuracy decay from temporal variance in the content stream

An interesting aspect of online supervised learning is that temporal variance in the content stream itself can reduce classification accuracy over time. This was demonstrated by Mourão et al. (2008) who identified three types of temporal effects: changes in class distribution, term distribution, and class similarity. Class distribution refers to the frequency with which different concepts are mentioned during different times. Term distribution refers to how different terms are used during different times to refer to a specific concept. Finally, class similarity may change over time, as different concepts become more or less related to each other over time.

These effects all reduce classification performance over time, by introducing new concepts, new features for existing concepts, or by moving decision boundaries between classes. Similar effects have been observed for much shorter time intervals in search query logs and in social media data collected during natural disasters. Kulkarni et al. (2011) showed how both search query terms and search hits relevant to specific search terms can vary greatly between days, indicating short-term presence of changes in class distribution and class similarity.

Figure 7.4 suggests that in rapid onset disasters, as much annotation work as possible should be performed as early as possible. However, if there is significant temporal variance in the content stream, classification performance will decay over time. If annotation effort is considered to be a fixed resource, significant temporal variance would make it necessary to distribute the effort more evenly over the duration of data collection. This aspect of supervised stream classification was investigated by Imran et al. (2014), who found that for rapid onset disasters in which data collection lasts up to a few weeks, term distributions can be considered relatively stable and accuracy does not drop substantially over the duration of the disaster. However, some message classes only appear in the stream during the response or recovery stages, thus if all classification is performed during the build-up or impact stages, no training data will be available for such

classes. While it was not part of the study, temporal variance is likely to play a greater role for work scheduling in longer lasting disasters such as conflict, disease outbreak or drought.

Furthermore, classification accuracy for future data can never be known with certainty, but all system using supervised learning need some way to communicate the estimated accuracy of the classifiers to end users, to know when no further training data needs to be provided. Traditionally this is done by putting aside some of the human-labelled training data to be used for model evaluation.

However, the temporal variance in the content stream implies an additional need for continuous monitoring of classifier accuracy, as classification performance is expected to degrade over time. Furthermore, a user should be able to get an indication of this degradation without continuously needing to annotate new evaluation data, in particular if model performance has already reached sufficient levels. Thus there is a need for metrics that can be calculated without availability of human-labelled data.

Possibly, this could be achieved by plotting over time the distribution of output labels as well as the distribution of associated confidence scores. As term distribution, class distribution and class similarity change over time, these distributions should change as well. The precise relationship between these distributions is however unclear, and more research is needed to understand how these or other metrics correlate with drops in classifying accuracy.

7.4 Conclusion

Content classification is a promising approach to help identify reports on online social media that relate to the information needs of specific decision makers. In addition, if both event detection and classifiers are of sufficient quality, they would together provide an estimate for the number of unique new incidents over time, which is a valuable quantitative indicator that it is so far not possible to track.

Although some systems have been proposed which include pre-trained static classifiers for social media content, the use of classifiers in disaster information management is very limited. The primary obstacles that hinder addition of such tools include that information needs may change between disasters, that the language in which future content is authored is unknown, and that classifiers trained on datasets collected in one disaster generalize poorly even to other disasters of the same type.

This chapter introduced online supervised learning as a potential way to handle disaster-specific classification needs and temporal variance in the content stream. A classification module was

built for the more comprehensive AIDR system and was integrated into CrisisTracker to provide classification of stories.

Experience with the system so far suggests that the new classification module currently performs fairly well in long-lasting disasters such as civil wars, but that its use is limited in short-lived rapid onset disasters, as extensive labelling effort must be carried out very early during such disasters for the system to be able to classify the bulk of the communication.

System use also revealed that additional development is needed before it becomes feasible to apply the system to identify uncommon types of reports. In addition, several directions for future research were proposed.

Chapter 8

Conclusion

This research has investigated how software tools can be used to sample and aggregate the collective sensory capacity of online social media users, to improve the situational awareness of decision makers in humanitarian disasters. The findings are primarily of interest to international organizations wishing to conduct real-time distance monitoring of events, and to the technical community that builds software tools targeting any of the primary stakeholders in humanitarian disaster response.

Previous research has shown that a large volume of social media communication contains information that would be valuable in decision making. However, the signal-to-noise ratio of the medium is too low and the rate of content production too high to allow effective monitoring without assistance from purpose-built software tools. Information management tools have so far been incapable of processing torrents of unstructured text, images and video to organize the information into a structure that reliably improves perception and comprehension of ongoing events.

Several concrete problems were identified in section 1.2 that needed to be addressed to enable practical use of online social media useful as an information source in disaster response. This chapter discusses the contributions and limitations of the research presented in this thesis against the problem statement, and suggests theoretical and practical directions for future work.

8.1 Contributions

8.1.1 Comprehensive documentation of stakeholders' roles and information needs

Information management in disaster response is highly complex, with a great number of stakeholders, each with their own multifaceted requirements for information on which they build their

situational awareness. Knowledge of both stakeholder roles and information needs is crucial throughout any information system design process, but in the disaster information management domain, such documentation of information needs has been missing (Verity 2011).

This lack of knowledge has had a significant impact on the design of related information management systems. For instance, even one of the most popular systems used for management of social media reports in this domain has been found to only partially support two of its core use cases: situational assessments and decision making (Morrow et al. 2011). Furthermore, statements regarding how proposed systems and workflows support information gathering and who is the intended user have been weak or missing entirely from much of previous research, and proposed systems have seen very limited adoption by practitioners.

Building on a recent workshop that mapped out the stakeholders in humanitarian disaster response (Verity 2013), section 2.4 covered five of the core groups: victims and on-site volunteers; public sector organizations; international organizations; online volunteers and the technical community; and mainstream media. For each of the groups, the section defined their overall role in the response effort, as well as the information needed and produced by that group.

This review compiles findings from previous research as well as two novel case studies into the first comprehensive overview of information needs in the disaster response domain. In addition to its breadth, to the author's knowledge the chapter also provides the first documentation of information needs of victims and on-site volunteers , as well as online volunteers and the volunteer technical community. Previous research studying any of these groups has been limited to documenting and discussing their activities and roles, without specifically discussing the information that members of these groups require to function in their roles.

This review now makes it easier for the disaster information management community to design, evaluate and compare information management solutions. Clear requirements will help researchers and practitioners target specific users and information types, as well as to compare existing practice against needs to find valuable areas of novel research. Several implications of these findings for the direction of future research are discussed in section 8.3.

Although the review has been performed with the intention of building tools for online social media monitoring, the documented roles and information needs are not specific to this information source and should be applicable also to disaster information management projects targeting other media, including offline workflows.

8.1.2 Intrinsic value cannot compensate for lack of payment in for-pay task markets

Crowdsourced human-based computation is a technique that has been successfully employed for many computational tasks that are easy to perform for humans, but challenging for computers. Many types of problems for which crowdsourcing have been used also have direct applications in the humanitarian domain, including image labelling, audio transcription, re-formatting of data, and information searches.

A large portion of crowdsourced work takes place on for-pay task markets, where workers are reimbursed financially for their time. Through such markets it is possible to enlist large workforces with short notice, which would be a desirable property during disaster response. However, crowdsourcing in the humanitarian domain has so far mainly been performed on a volunteer basis, and it is not desirable to introduce high operational costs, nor is it clear if there are qualitative differences between for-pay and charitable work. A study, presented in chapter 4, was therefore conducted to understand if workers on for-pay task markets would be willing to donate their time for charity, as well as how variations in payment would affect quality of work.

Contrary to limited previous research (Chandler and Kapelner 2013), the study concluded that payment alone, not the intrinsic value of the work, affected uptake and productivity of workers. Not-for-pay work completed at rates far too low to be of any practical use in time-pressed disaster situations.

These findings imply that for-pay task markets are unlikely to be well-suited for use in human-based computation for humanitarian purposes. Consequently, the attention in later chapters was shifted towards not-for-pay groups such as the Stand-By Task Force and smaller but highly motivated analyst teams. A yet unexplored application of for-pay task markets is during deployments with very limited duration when sufficient payment can be provided, for instance during the critical first hours following a major natural disasters. However, if future research pursues this direction, or in other ways introduces financial rewards, significant care needs to be taken to not cause long-term persistent negative effects on work quality beyond the point when money is no longer offered, similar to what occurred in an experiment by Gneezy and Rustichini (2000a).

We also found that increasing the intrinsic value of the work significantly increased worker accuracy when the work paid poorly or not at all. Though the effect size was substantial in our experiment, further research would be needed before any predictive statements in this regard can be made. These findings, presented in section 4.6.4, are in contrast to previous work by Chandler and Kapelner (2013) who presented weak evidence suggesting that worker accuracy was unaffected by intrinsic motivation. At higher payment levels however this effect disappeared, in agreement with previously proposed theory (Gneezy and Rustichini 2000b),

which argued that extrinsic motivators can override the intrinsic value of work and eliminate its positive effects.

The finding that meaningful work results in high quality output is encouraging for the humanitarian community, as it may permit crowdsourcing of more complex information seeking or reasoning tasks, for which it may be difficult to verify the correctness of answers from crowdsourcing workers. Further research is needed to investigate if intrinsic motivation itself under some conditions can be a sufficient replacement for commonly used quality control techniques such as averaging of redundant answers, or gold standard examples.

8.1.3 With current technologies, effective social media monitoring requires hybrid workflows

Section 2.3 identified substantial evidence in literature that information corresponding to many of the information needs documented in section 2.4 is exchanged through online social media during ongoing humanitarian disasters. Effective real-time monitoring of public social media communication thus has the potential to tap into a vast sensor and information filtering network, consisting of thousands or even millions of users distributed throughout the disaster area and the rest of the world.

However, practical attempts to monitor this information source have so far only been possible during short periods of time or for highly specific purposes known well in advance of the monitored event . This is because the new medium poses several new challenges not seen in traditional document corpuses. In particular, social media corpuses consist of very large numbers of documents, have short individual document length, very low signal to noise ratio and high redundancy . These and other challenges are discussed in sections 2.3.1-2.3.2.

The state of the art approach for social media monitoring during humanitarian disasters has been to rely almost entirely on crowdsourced human-based computation, primarily using the Ushahidi platform[1]. This technique has been successful due to its exceptional flexibility in handling novel forms of data collection and information extraction. However, the approach is slow, workers are prone to burn-out, and as content producers can number in the millions, it is not reasonable to expect that content curating crowds can ever manually inspect all new content (Meier 2013). Furthermore, available software tools have provided little workflow support, resulting in failures to scale efforts to handle true disasters or to sustain them during prolonged crises (Meier 2012). The general consensus among the practitioner sources interviewed during the 2013 field study was that reliable real-time monitoring has never been achieved.

[1] http://www.ushahidi.com/

Several techniques for scalable machine-based computation have therefore been applied in attempts to filter and summarize social media streams into information products with greater utility in decision making (e.g. (Kumar et al. 2011; Yin et al. 2012; Abel et al. 2012; Jadhav et al. 2010)). However, as explained in section 3.1, it has become apparent that information extraction algorithms are not yet ready to fully replace human curators in this application domain, due to the quantity, brevity, informality and non-English nature of much of the communication.

Because of the still limited capabilities of machine-based processing techniques, human-based computation will remain a necessary component of social media monitoring in disasters in the foreseeable future. However, because human-based computation is a limited resource with comparably low throughput, successful monitoring is most likely to be achieved using hybrid workflows, which integrate the complementary strengths of each technique.

By identifying and automating expensive sub-tasks in otherwise manual workflows, it will become possible to apply human cognition to solve complex computational problems at increasingly larger scales. As technology progresses and analysis processes become more standardized, focused research effort can be invested to automate increasingly greater portions of the workflow. This will eventually allow system users to fully shift their attention towards the higher cognitive aspects of situational awareness building – comprehension and projection – which is further discussed in section 8.3.1.

Workflow support has been the primary focus of the technical research presented in this thesis, and the proposed solutions have been implemented and evaluated in the open-source CrisisTracker system. Specific findings and their significance are discussed below.

8.1.4 Clustered social media feeds can improve the situational awareness of international humanitarian organizations

Previous research studying social media monitoring for humanitarian purposes has encountered several obstacles that have prevented scalable high-quality processing of social media communication during disasters (Gao, Barbier, and Goolsby 2011; Morrow et al. 2011; Verity 2011). Many communication platforms, in particular Twitter, rely on content duplication to propagate information between users. Effective message de-duplication strategies are thus needed to avoid that the same information is processed multiple times, wasting precious human resources. Messages are also so short that a comprehensive understanding of an event often can only be reached through inspection of multiple messages, and techniques are needed that can group related content into more comprehensive stories. Furthermore, as the number of unique reports far surpasses that which can be manually processed or consumed, techniques are needed that direct curators towards information that has a high probability to be beneficial to decision makers. This is true both in terms of ranking of unique stories, and within-story sampling or

summarization. Solutions to all these problems are needed to enable new event detection, which would help drive attention towards new developments in a situation that may require intervention. Finally, to be of practical utility, any solutions must be capable of processing information in local languages spoken by disaster-affected communities.

This thesis proposed addressing these challenges through automated document clustering. The Locality Sensitive Hashing (LSH) algorithm (Charikar 2002) was identified in section 5.2 as an online, high-throughput and language-agnostic solution, which was further extended in sections 5.3.1 and 6.2.2 to improve cluster recall in the application context. Techniques for ranking and summarizing clusters were proposed in sections 6.2.3 and 7.2.2. The algorithms were implemented in the open-source CrisisTracker[2] system, which is freely available for download. The integrated solutions in CrisisTracker were evaluated against alternative information gathering practices already used by practitioners, through a field study focused on the 2011 Syrian Civil War, to make comparisons in terms of timeliness, detection reliability and richness.

Experienced disaster response professionals who used the system concluded that CrisisTracker effectively aggregates information from large numbers of individually sparse reports into rich comprehensive stories that well describe unfolding events. Trained analysts reported that CrisisTracker's stories helped them better detect and comprehend both high-level geo-political connections between events (section 6.4.3), and highly localized relationships between different pieces of evidence for a single event (section 6.4.4). By bringing together related messages, users also claimed the system helped them conduct credibility assessments, by connecting evidence spread across different reports and by showing how individual independent accounts either corroborate or conflict each other. This was of particular value in a conflict setting, when individual accounts were often seen as partial to either side in the conflict.

Evaluation shows that the system detects new events on par with or earlier than mainstream media, when filtering thresholds are still set to avoid information overload (section 6.4.4). In terms of timeliness, CrisisTracker was shown to reliably detect the most impactful events within 30 minutes of the first tweet being posted. During the studied conflict, participating analysts who actively monitored the crisis in parallel using established techniques[3] claimed that they often detected stories in CrisisTracker hours and sometimes even days before they saw coverage in traditional news media. Application of the methods proposed in this thesis thus enable better allocation of complementary and more costly information gathering resources (e.g. high-resolution satellite imagery or field teams), better informed decision making, and earlier intervention. In the theoretical framework of situational awareness (see section 2.2), these findings correspond to contributions at level 1 (perception) and level 2 (comprehension), while contributions at level 3

[2] http://github.com/JakobRogstadius/CrisisTracker/

[3] Manual and automated monitoring of mainstream media, manual monitoring of social media, and trusted contacts on the ground.

(projection and prediction) are indirect and directly dependent on the experience and training of the system user.

Both feedback during the field study and comparison against the documented information needs indicate that the content ranking algorithms implemented in CrisisTracker highlight reports that best match the information needs of international response organizations. Lessons learned were therefore discussed in section 6.5.3 regarding how to integrate social media monitoring tools into organizational workflows, including recommendations to give analysts rather than decision makers access to the tools, as well as to establish a new crowd director role.

8.1.5 Clustering and supervised learning help scale human-based curation of social media feeds

No single content ranking algorithm will successfully match users' information needs of every situation, and information management tools must therefore provide additional search and filtering capabilities. In the disaster response domain, filtering may be used for instance to single out reports relating to logistics, food, population movements or other humanitarian needs sectors, to a particular geographic region, or to a specific affected demographic population group. Such filtering requires accurate extraction of meta-data, to turn unstructured text into structured data. Extensive meta-data extraction is also required to accurately compute temporal trends, variations and geographical patterns in the collected knowledge.

The prevalent method for such information extraction has been crowdsourced human-based annotation of individual pieces of information (Meier 2013), whether they are images, videos, SMS, or social media posts. However, available tools for supporting this workflow have lacked sampling strategies necessary to prioritize work items, which is problematic as the inflow of unprocessed reports typically far surpasses the collective processing capacity. Because of this, end-users have questioned the representativeness of the underlying data for derived information products (Morrow et al. 2011).

Rather than annotating a random subset of reports, section 5.5.4 showed how clustering of social media messages can provide ranking metrics which can direct workers towards content that is more likely to at a given time maximize the gain in situational awareness. The metric used is the extent to which a large subset of social media users, who are closely associated with the main disaster, are showing interest in a specific story at a given time. Metrics were also defined in section 7.2.2 to summarize stories through a small, representative and diverse sample of messages.

Clustering also has the additional benefit of greatly reducing the total workload, through de-duplication and grouping of related content. By assigning stories rather than individual messages

as work items, meta-data that is assigned to a story by a human curator automatically applies to all messages within the group, including updates that are authored after the task was completed. Clustering thus makes the total workload scale with the total volume of information, proportional to disaster scale and complexity, rather than the total volume of communication, proportional to the number of sources that report on the disaster. Section 6.4.1 presented a quantitative analysis of this effect, suggesting that only around 1 in 1000 stories were reported by more than a handful of sources. By focusing on these highly reported events, content curators were able to process over 600,000 tweets in around 130 total work hours; a substantial improvement over previous efforts that used similar processing tasks.

To further increase scalability, the clustering functionality was complemented with the development of a processing architecture for online supervised classification of message streams (chapter 7). This system is capable of learning user-defined classification schemes from human curators and to apply them on future content. Integration of supervised learning addresses the fundamental issue of increasing the total work output of a given crowd to keep up with message inflow rates several orders of magnitude greater than what can be processed using human-based computation alone. Further development is however needed to improve the classification system's utility for sparse classes and during rapid onset disasters.

8.2 Limitations

As noted in the introduction, the research has focused exclusively on increasing the utility of a single information channel; online social media. Disaster response is complex, and in practice this channel is merely one of several from which responders collect and aggregate information. This means that the proposed techniques are also only useful during disasters in communities that actively use social media, effectively making them inapplicable to parts of the developing world that have yet to experience a boom in internet connectivity, or in markets where the most popular social media service providers restrict public access to the published content. Utility naturally suffers when internet connectivity is temporarily disabled locally, though use of the system does suggest that reports still reach the public indirectly, although more slowly, in smaller numbers, and never with first-hand accounts.

At the time of writing, independent deployments of CrisisTracker have been set up by a number of research groups, media organizations and conflict monitoring organizations. To the author's knowledge, the deployments have however only made use of the fully automated features in the system, with little manual curation taking place. This implies that the meta-data extraction workflow, though similar to that in the popular Ushahidi system, still needs to be improved. To create substantial benefits over existing tools, a new system likely needs to both provide extraction of additional meta-data beyond time, keywords and labels, in accordance with the

documented information needs, as well as meaningful visualizations that make use of that meta-data to aid in predictive analysis and detection of high-level patterns and trends.

That such features have not been implemented CrisisTracker is in great part due to that information needs were poorly understood at the onset of the research and that they only gradually became apparent as additional research was published by external sources. Meta-data types that were identified as important to decision makers but for which no good machine-based extraction technique yet exists are: report type, involved entities, quantitative metrics of humanitarian impact, geographic location, and vulnerable groups most strongly affected. Availability of these meta-data types would additionally make it possible to support local stakeholders such as victims, volunteers and public sector organizations, if it is also possible to develop suitable automated information ranking metrics, possibly through use of community detection algorithms.

It is also important to keep in mind that CrisisTracker has no explicit support for report verification or rumour detection. While clustering of related content helps users assess the credibility of individual reports and while peer-correcting behaviour has been observed on Twitter (The Guardian 2011), it remains important that analysts rely on techniques such as cross-comparison of evidence from different independent sources to avoid false leads. Social media should not be used as the only information source for critical decisions.

Additional research is also needed to validate how well the proposed information management techniques perform in disasters of types other than conflict and civil unrest. System performance during events that last only for a few days can be assumed to be worse than during long-lasting or slow onset disasters, as both the clustering algorithm and the learning classifiers require some training time before they reach acceptable output quality. CrisisTracker's clustering algorithm has also been observed to break down when data is scarce, which happens during highly localized or less impactful events or in geographic areas where social media use is very limited. Furthermore, as processing is meant to be performed in real-time, worker recruitment can be an additional bottle-neck as it can be challenging to enlist large work forces with short notice.

The clustering algorithm used in CrisisTracker also has limitations even under the most favourable conditions. First, there is no guarantee that messages within a cluster are semantically related, as the algorithm relies only on shallow bag-of-word techniques, though such techniques were selected intentionally due to their language independence. Furthermore, the algorithm's recall is sub-optimal and many duplicate stories are produced for events that are covered from many angles. In addition, the proposed second-order clustering algorithm is iterative rather than online, which makes it computationally expensive compared to the LSH algorithm that is used for the first-order clustering. Furthermore, both algorithms have many tuning parameters that require in-depth knowledge of their inner workings to adjust for variations in the data stream.

Despite these limitations, the value of online clustering for social media monitoring is clearly evident and future research should strive to identify alternative more robust algorithms.

8.3 Future work

8.3.1 Support transfer of higher level situational awareness

From a conceptual point of view, future research should investigate ways to help users collaboratively reach increasingly higher levels of situational awareness. CrisisTracker currently improves the flow of situational reports from the general public to users of the system. This process helps responders get an improved perception of the environment (level 1 situational awareness). If a system deployment is publicly accessible and used by many stakeholder groups, it can also contribute to a shared perception, which can reduce confusion and misunderstandings in communication. Future research should investigate how disaster information management systems can better support sharing of higher level situational awareness, as well as investigative queries triggered by insights.

8.3.1.1 Comprehension and projection

According to situational awareness theory (Endsley 2000), operators who possess significant local awareness, personal experience and professional training are better at interpreting and reasoning about perceived facts. This means that compared to laymen, they can reach a deeper level of understanding, which is often required to reach level 2 and 3 situational awareness (comprehension and projection). According to subjective feedback from users, CrisisTracker currently offers some degree of support for this process, by presenting related information together, so that patterns, event timelines and causal relationships can be better understood.

However, as users have varying backgrounds and levels of training, they comprehend information differently and are able to make different projections. In particular, as documented in chapter 2, most responders in humanitarian disasters are not trained professionals, but rather regular citizens who rise to face needs in their local community. These individuals do not have the experience required to comprehend complex patterns and to anticipate future hazards or actions of other responders.

Future disaster information management systems should therefore support and encourage users to feed their reflections, conclusions and hypotheses back into the system, and to let others build on those insights, as this would enable transfer of comprehension and projection between individuals and between stakeholder groups. While regular citizens can already provide experienced responders with access to raw information, such features would introduce an additional flow in

the opposite direction, to give inexperienced citizens greater access to the expert knowledge currently available only to trained professionals.

One way to support this in practice could be to introduce an additional user interface layer of collaborative reasoning. For instance, a traditional wiki or discussion forum could be extended with support for linking forum posts and wiki articles to stories or events in the system, and to annotate them with the same meta-data structure as the collected stories. This would enable visualization of reflections and insights alongside the reports, both as comments on individual stories, on maps and timelines, and in search results.

8.3.1.2 Investigation

Neither the current functionality in CrisisTracker nor the proposed features for shared comprehension and projection would actively drive the generation of new information. However, with increased reasoning and collective access to insights comes awareness of unverified claims or a desire for information that is not yet available.

Systems should therefore allow users to make information requests, to which other users of the system can submit evidence. Requests for verification likely need to separate evidence for or against the claim. When sufficient evidence has been collected and either the community or original requester deems that a conclusive answer has been reached, the new knowledge should be merged into the same knowledge base from which the original question was posed. Research is needed to find ways of integrating such knowledge in practice, as well as to find reward mechanisms that generate timely and high quality answers to questions posed in the system.

Unlike traditional news media, online social media in theory makes it possible for anyone to directly interact with information sources. This opportunity can be of tremendous value for verification of new claims and for emergent coordination among volunteers. However, there are significant risks associated with revealing the identity of victims and in conflict situations, including exposing sources as targets of violence. Future work should identify ways to facilitate direct communication between original sources and information analysts, while minimizing the risk to both parties.

8.3.2 Allocate processing resources by information need, not information availability

CrisisTracker has been built to make it easier to detect reports that mention new events, with greater priority given to clusters of reports that have seen greater interest from social media

users strongly associated with the disaster. Its application of clustering and scalable metadata extraction has been shown to prevent information overload and increase situational awareness. However, by building on these proposed information processing techniques, future systems should strive to pre-process and visualize information in ways that more closely match the specific information needs described in chapter 2, which in turn would let them better support decision making.

Stakeholders with high-level interests, primarily international organizations with a coordinating role, would be best supported if presented with summaries in the form of patterns and trends; maps depicting the humanitarian situation in different affected regions, time trends of the number of people affected by hazards and reached by interventions, comparisons of needs by humanitarian sector, etc. Other stakeholders would be better supported by a structured event model. Conflict analysts are interested in tracking the (co-)involvement of different entities in different events, and responders who coordinate intervention or themselves actively intervene likely want access to systems in which they can track the status of specific issues. The main point is that both aggregate views and event models are examples of fitting available information into a structured model, rather than attempting to infer a structure from available information.

The proposed method of information processing uses the steps: collect, cluster, index, sample, process, visualize. First, automatically collect reports from social media and perform message de-duplication using the techniques proposed in chapter 5 and 6. This process groups together messages that contain equivalent and/or closely related information, and identifies within-group messages that are both informative and diverse. If desired, automated summarization algorithms can be explored to further improve the information coverage of the summary, but additional research may be required to understand how such techniques can be employed in a streaming context.

After de-duplication, the number of clusters can be assumed to still be several orders of magnitude greater than what any group of humans can realistically process. The next step is therefore to develop classifiers that can be used to index the information, or to sort messages into suitable bins corresponding to some of the dimensions of information needs described in chapter 2. For the set of stakeholders that was reviewed, the dimensions of this structure are: report type, humanitarian sector, time, geographic location, vulnerable groups, named entities, intervention status and quantifiers of humanitarian impact. It is unlikely that accurate classifiers can be developed for every dimension, thus the goal is to structure the information such that it becomes possible to select a representative sample from each bin, which can then be further processed by human assessors.

Crucially, human-based processing resources should be allocated based on humanitarian needs, rather than availability of information. For instance, reporting frequency will largely correspond to population density, but areas will less population may be in greater humanitarian need and

should receive greater attention. Similar arguments of bias can be made also for other dimensions of information, such as sectors or report types. Automation should be used to pre-process and re-sample the information stream to avoid skewed distributions in reporting frequency that arise from both the uneven public interest in different topics and the ease of reporting certain types of events.

By mapping reports into an information structure, it also becomes possible to identify combinations of meta-data for which few or no reports are available. These scarce regions of the information space are prime candidates for information seeking tasks.

Finally, whether the goal is to construct a relational event model or to populate the cells of an aggregate data cube, the original reports from which the information was inferred should be retained as evidence that has an associative relationship with the knowledge. By keeping reports as linked evidence rather than as the main information carriers in the system, it becomes straightforward to integrate different parallel source media into the system, which has been difficult to do in CrisisTracker.

8.3.3 Study poorly understood roles of information in disasters

The information space in humanitarian disasters is incredibly complex and there is a clear need to further document the information needs and decision making processes of the stakeholders not covered in this thesis.

Furthermore, it would be greatly beneficial to document how information flows between different stakeholders, in addition to how information is produced and consumed. In this context, an information flow is when knowledge is produced by one stakeholder group and then shared with others to improve their situational awareness or decision making. Several tools, like CrisisTracker, currently support the flow of information from the general public to public sector and international organizations. However, far fewer efforts are focused on supporting flow of information in the opposite direction, to make high-resolution event models, interpretations and projections available to victims and on-site volunteers. For more information on why this is important, see (Rogstadius et al. 2013a). By further mapping disaster information flows, it would become possible to identify both strongly supported flows where there is little need to allocate additional resources, and weakly supported flows, where information that would be useful is currently generated but inaccessible.

Researchers and system developers could also benefit from a deeper understanding of when different stakeholders are present, in terms of crisis scale, community resilience or development level, violence vs. natural disasters and slow- vs. rapid onset disasters. This would help map environmental and information constraints to stakeholder needs.

There is also a need to better understand the effects of making high-resolution disaster-related information publicly accessible to victims and volunteers, in particular in conflict environments. For instance, both researchers and practitioners should strive to understand what information, when made public, drives actions that have a positive influence on the humanitarian situation, as well as what information may lead to actions that have a negative impact, for instance by fuelling anger or even revenge. Documenting such processes would likely provide directions in which open-access information management systems can improve the safety of regular citizens or help them contribute to peace and recovery.

Specifically in conflict monitoring situations, there is also a need to establish best practices for how information should be processed and stored to avoid contaminating evidence that could become useful in post-war prosecution of war criminals. CrisisTracker already provides a step in the right direction compared to some other systems, by guaranteeing that reports are kept in their original format and by keeping meta-data and higher-level structure as a separate layer of annotations, but further steps could be taken for instance by aggregating evidence from multiple media, indexing reports containing first-hand accounts, maintaining (protected) lists of first-hand witnesses, and by fully separating event models from raw evidence.

8.3.4 Develop ethical guidelines for humanitarian information management

Technologies that are capable of enabling new forms of open source intelligence can vastly change the relationship between information value and information risk. An example is that a single public tweet may pose minimal risk to its author, but the risk changes substantially if a longitudinal analysis that examines thousands of users' message history identifies the person as the sole reliable source of evidence documenting long-term atrocities in an area. While closely related to the topics of this thesis, such ethical concerns are an area of research that has only been touched upon briefly herein.

Ethical handling of information is particularly important for analysis that is made public, partially or entirely, and in particular during conflict or other events when human intervention is a real and direct threat to vulnerable groups. While disaster response organizations have traditionally kept information primarily for internal use, recent and future methods of data sharing have great potential benefits and may even be a necessity for some forms of crowdsourced processing. The established practice of blanket confidentiality is thus becoming far more nuanced. This was illustrated for instance by a crisis map set up to track events during the 2011 Libyan Civil War, which was initially set up for internal use only, but later configured to release reports publicly with a 24 hour delay, to benefit other organizations while limiting its utility in planning armed attacks (SBTF & UN OCHA 2011).

Humanitarian intervention generally operates on a do-no-harm basis and ethical principles are established for intervening humanitarian work. However, while crisis mapping and social media monitoring have been established as valuable and powerful tools, their impact and risks are not yet fully understood and ethics is a very active topic of debate among practitioners and researchers. Therefore, it would be of value to the volunteer technical community if the Red Cross' Code of Conduct in Disaster Relief (ICRC 1994), the de-facto standard in ethical disaster response, could be translated into generally applicable guidelines for information management, which can be directly applied to ensure that software tools are designed to be ethically sound. This would be of particular value if new tools or communication platforms are developed to support self-help efforts and spontaneous volunteering in affected communities.

Bibliography

Abel, Fabian et al. (2012). "Semantics + filtering + search = twitcident. exploring information in social web streams". In: *Proceedings of the 23rd ACM conference on Hypertext and social media*. New York: ACM, pp. 285–294.

Ahn, Luis von and Laura Dabbish (2004). "Labeling Images with a Computer Game". In: *Proc CHI'06*. Vienna, pp. 319–326.

Beenen, Gerard et al. (2004). "Using social psychology to motivate contributions to online communities". In: *Proc. CSCW '04*. New York: ACM Press, pp. 212–221.

Bernstein, Michael S. et al. (2010). "Eddi: interactive topic-based browsing of social status streams". In: *Proc. UIST '10*. New York: ACM Press, pp. 303–312.

Best, Clive et al. (2005). *Europe Media Monitor - System Description*. Tech. rep.

Bhuiyan, Serajul I. (2011). "Social Media and Its Effectiveness in the Political Reform Movement in Egypt". In: *Middle East Media Educator* 1.1, pp. 14–20.

Brattberg, Erik (2013). *The case for US military response during international disasters*. http://thehill.com/blogs/congress-blog/foreign-policy/190954-the-case-for-us-military-response-during-international. Accessed: 2013 Dec 14.

Brennan, M.A., Rosemary V. Barnett, and Courtney G. Flint (2005). "Community Volunteers: The Front Line of Disaster Response". In: *International journal of volunteer administration* 24.4, pp. 52–56.

Chandler, Dana and Adam Kapelner (2013). "Breaking monotony with meaning: Motivation in crowdsourcing markets". In: *Journal of Economic Behavior & Organization* 90, pp. 123–133.

Charikar, Moses S. (2002). "Similarity estimation techniques from rounding algorithms". In: *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*. New York: ACM, pp. 380–388.

Chowdhury, Abdur (2011). *Global pulse*. https://blog.twitter.com/2011/global-pulse. Accessed: 2013 Sep 18.

Cisco (2012). *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2011–2016*. http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html. Accessed: 2013 Jan 8.

Cooper, Seth et al. (2010). "Predicting protein structures with a multiplayer online game". In: *Nature* 466.7307, pp. 756–760.

Cosley, Dan et al. (2005). "How oversight improves member-maintained communities". In: *Proc. CHI '05*. New York: ACM Press, pp. 11–20.

Deci, Edward (1975). *Intrinsic Motivation*. New York: Plenum Publishing Company Limited.

Dynes, Russell R. (1994). "Community emergency planning: False assumptions and inappropriate analogies". In: *International Journal of Mass Emergencies and Disasters* 12.2, pp. 141–158.

Dynes, Russell .R., E.L. Quarantelli, and Dennis Wenger (1990). *Individual and organizational response to the 1985 earthquake in Mexico City*. Tech. rep.

Emergency Management Australia (2006). *Hazards, disasters and your community, 7th ed.* Tech. rep.

Endsley, Mica (2000). "Theoretical underpinnings of situation awareness: A critical review". In: *Situation Awareness: Analysis and Measurement*. Routledge, pp. 3–32.

Etzioni, Amitai (1971). *Modern Organizations*. Engelwood Cliffs, New Jersey: Prentice-Hall.

Faris, David (2008). "Revolutions Without Revolutionaries? Network Theory, Facebook, and the Egyptian Blogosphere". In: *Arab Media & Society*, pp. 1–11.

Finin, Tim et al. (2010). "Annotating named entities in Twitter data with crowdsourcing". In: *Proc. CSLDAMT '10*. Stroudsburg, PA: ACM Press, pp. 80–88.

Gao, Huiji, Geoffery Barbier, and Rebecca Goolsby (2011). "Harnessing the Crowdsourcing Power of Social Media for Disaster Relief". In: *IEEE Intelligent Systems* 26.3, pp. 10–14.

Ghannam, Jeffrey (2011). *Social Media in the Arab World: Leading up to the Uprisings of 2011*. Tech. rep. Washington.

Gibbons, Robert (1997). "Incentives and Careers in Organizations". In: *Advances in Economic Theory and Econometrics* II. Ed. by D. Kreps and K. Wallis.

Gneezy, Uri and Aldo Rustichini (2000a). "A fine is a price". In: *Journal of Legal Studies* 29.1, pp. 1–17.

– (2000b). "Pay enough or don't pay at all". In: *The Quarterly Journal of Economics* 115.3. Ed. by Robert J. Barro et al., pp. 791–810.

Gonzalez, Michael M. (2005). *Citizen Involvement in Disaster Management*. Tech. rep. Monterey, CA.

Graham, Mark (2012). *What can Twitter tell us about Hurricane Sandy flooding? Visualised*. http://www.guardian.co.uk/news/datablog/2012/oct/31/twitter-sandy-flooding.

Graham, Mark, Ate Poorhuis, and Matthew Zook (2012). *Digital trails of the UK floods - how well do tweets match observations?* http://www.guardian.co.uk/news/datablog/2012/nov/28/data-shadows-twitter-uk-floods-mapped.

Gralla, Erica, Jarrod Goentzel, and Bartel Van de Valle (2013). *Report from the workshop on field-based decision makers' information needs in sudden onset disasters*. Tech. rep.

Green III, Walter G. (2003). "Freelance Response to the Site - Medical Staff Option of Choice?" In: *The AAMA Executive*, pp. 1–16.

Harb, Zahera (2011). "Arab Revolutions and the Social Media Effect". In: *M/C Journal* 14.2.

Harper, F. Maxwell et al. (2008). "Predictors of answer quality in online Q&A sites". In: *Proc. CHI '08*. New York: ACM Press, pp. 865–874.

Harvard Humanitarian Initiative (2011). *Disaster Relief 2.0: The Future of Information Sharing in Humanitarian Emergencies*. Tech. rep. Washington, D.C. and Berkshire, UK.

– (2012). *Uchaguzi: A Case Study*. http://reliefweb.int/sites/reliefweb.int/files/resources/uchaguzi-121024131001-phpapp02.pdf.

Heer, Jeffrey and Michael Bostock (2010). "Crowdsourcing graphical perception: using Mechanical Turk to assess visualization design". In: *Proc. CHI '10*. New York: ACM Press, pp. 203–212.

Hofmann, Charles-Antoine and Laura Hudson (2009). "Military responses to natural disasters: last resort or inevitable trend?" In: *Humanitarian Exchange Magazine* 44.

Horton, John J., David G. Rand, and Richard J. Zeckhauser (2011). "The online laboratory: Conducting experiments in a real labor market". In: *Experimental Economics* 13.3, pp. 399–425.

Houle, David (2009). *Tiananmen Square and Technology*. http://www.evolutionshift.com/blog/2009/06/03/tiananmen-square-and-technology/. Accessed: 2013 Jan 8.

Howard, Philip N. (2011). *The Arab Spring's Cascading Effects*. http://www.psmag.com/politics/the-cascading-effects-of-the-arab-spring-28575/. Accessed: 2013 Jan 8.

Hsieh, Hsiu-Fang and Sarah E. Shannon (2005). "Three Approaches to Qualitative Content Analysis". In: *Qualitative Health Research* 15.9, pp. 1277–1288.

Hubbard, R. Glenn and Darius Palia (1995). "Executive pay and performance: Evidence from the US banking industry". In: *Journal of Financial Economics* 39.1, pp. 105–130.

Hughes, Amanda Lee and Leysia Palen (2009). "Twitter Adoption and Use in Mass Convergence and Emergency Events". In: *Proceedings of the 6th International ISCRAM Conference*. Gothenburg, Sweden: J. Landgren and S. Jul, eds.

Humanitarian Coalition (2013). *What is a humanitarian crisis?* http://humanitariancoalition.ca/info-portal/factsheets/what-is-a-humanitarian-crisis. Accessed: 2013 Dec.

ICRC (1994). *Code of Conduct for the International Red Cross and Red Crescent Movement and Non-Governmental Organizations (NGOs) in Disaster Relief*. Tech. rep.

IFRC (2000). *Disaster Emergency Needs Assessment*. Tech. rep.

– (2005). *World Disasters Report: Focus on Information in Disasters*. Tech. rep.

– (2013a). *Types of disasters: Definition of hazard*. http://www.ifrc.org/en/what-we-do/disaster-management/about-disasters/definition-of-hazard/. Accessed: 2013 Dec.

– (2013b). *What is a disaster?* http://www.ifrc.org/en/what-we-do/disaster-management/about-disasters/what-is-a-disaster/. Accessed: 2013 Dec.

IFRC (2013c). *World Disasters Report 2013*. Tech. rep.

Imran, Muhammad, Ioanna Lykourentzou, and Carlos Castillo (2013). "Engineering Crowdsourced Stream Processing Systems". In: *arXiv preprint*. ISBN: arXiv:1310.5463.

Imran, Muhammad et al. (2013a). "Extracting information nuggets from disaster-related messages in social media". In: *Proceedings of the 10th International ISCRAM Conference*. Baden-Baden, pp. 1–10.

– (2013b). "Practical extraction of disaster-relevant information from social media". In: *Proc. WWW '13*. Geneva: ACM Press, pp. 1021–1024.

Imran, Muhammad et al. (2014). "Coordinating Human and Machine Intelligence to Classify Microblog Communications in Crises". In: *Proceedings of ISCRAM 2014*. University Park, PA.

Inter-Agency Standing Committee (2012). *Multi-Cluster/Sector Intial Rapid Assessment (MIRA) - Provisional Version March 2012*. Tech. rep.

Internet World Stats (2012). *Internet Usage Statistics*. http://www.internetworldstats.com/stats.htm. Accessed: 2013 Jan 8.

Ipeirotis, Panagiotis G. (2010). "Demographics of Mechanical Turk". In: *New York University Working Paper No. CEDER-10-01*.

Jadhav, Ashutosh et al. (2010). *Twitris 2.0 : Semantically Empowered System for Understanding Perceptions From Social Data*. Tech. rep.

Jaeger, Paul T. et al. (2006). "The 2004 and 2005 Gulf Coast Hurricanes: Evolving Roles and Lessons Learned for Public Libraries in Disaster Preparedness and Community Services". In: *Public Library Quarterly* 25.3-4, pp. 199–214.

Jakobsen, Peter Viggo (2000). "Focus on the CNN Effect Misses the Point: The Real Media Impact on Conflict Management is Invisible And Indirect". In: *Journal of Peace Research* 37.2, pp. 131–143.

Java, Akshay et al. (2007). "Why we twitter: understanding microblogging usage and communities". In: *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*. New York: ACM, pp. 56–65.

Kaufmann, Nicolas, Thimo Schulze, and Daniel Veit (2011). "More than fun and money. worker motivation in crowdsourcing–a study on mechanical turk". In: *Proceedings of the Seventeenth Americas Conference on Information Systems*. Detroit.

Kittur, Aniket, Ed H. Chi, and Bongwon Suh (2008). "Crowdsourcing user studies with Mechanical Turk". In: *Proc. CHI '08*. New York: ACM Press, pp. 453–456.

Konkel, Frank (2013). *Tweets give USGS early warning on earthquakes*. http://fcw.com/articles/2013/02/06/twitter-earthquake.aspx. Accessed: 2013 Dec 5.

Kulkarni, Anagha et al. (2011). "Understanding temporal query dynamics". In: *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM Press, pp. 167–176.

Kumar, Shamanth et al. (2011). "TweetTracker: An Analysis Tool for Humanitarian and Disaster Relief". In: *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. Association for the Advancement of Artificial, pp. 661–662.

Kuznetsov, Stacey (2006). "Motivations of contributors to Wikipedia". In: *ACM SIGCAS Computers and Society* 36.2.

Lambert, Laura (2005). *The Internet: A Historical Encyclopedia*. Ed. by Hilary W. Poole, Leah Hoffmann, and Nicole Cohen Solomon. Santa Barbara: ABC-CLIO, Inc.

Lazear, Edward P. (2000). "Performance, Pay and Productivity". In: *American Economic Review* 50.5, pp. 1346–1361.

Lowe, Seana and Alice Fothergill (2003). "A need to help: Emergent volunteer behavior after September 11th". In: *Beyond September 11th: An Account of Post-Disaster Research*, pp. 293–314.

Mason, Winter and Duncan J. Watts (2009). "Financial incentives and the "performance of crowds"". In: *ACM SIGKDD Explorations Newsletter* 11.2, pp. 100–108.

McCarthy, John D., Clark McPhail, and Jackie Smith (1996). "Images of Protest: Dimensions of Selection Bias in Media Coverage of Washington Demonstrations, 1982 and 1991". In: *American Sociological Review* 61.3, pp. 478–499.

McNaboe, Christopher (2013). *Program Associate, The Carter Center*. Interview.

Meier, Patrick (2011). "New information technologies and their impact on the humanitarian sector". In: *International Review of the Red Cross* 93.884, pp. 1239–1263.

– (2012). *Some Thoughts on Real-Time Awareness for Tech@State*. http://irevolution.net/2012/02/02/real-time-awareness/. Accessed: 2012 Nov.

– (2013). *Humanitarianism in the Network Age: Groundbreaking Study*. http://irevolution.net/2013/04/09/humanitarianism-network-age/. Accessed: 2014 Mar 20.

Morrow, nathan et al. (2011). *Independent Evaluation of the Ushahidi Haiti Project*. Tech. rep.

Mourão, Fernando et al. (2008). "Understanding temporal aspects in document classification". In: *Proceedings of the 2008 International Conference on Web Search and Data Mining*. ACM Press, pp. 159–170.

Nofi, Albert A. (2000). *Defining and measuring shared situational awareness*. Tech. rep. Alexandria, Virginia.

Palen, Leysia and Sophia B. Liu (2007). "Citizen Communication in Crisis: Anticipating a Future of ICT-Supported Public Participation". In: *Proc. CHI 2007*. San Jose, California, USA: ACM, pp. 727–736.

Pan American Health Organization (2000). *Natural Disasters: Protecting the Public's Health*. Tech. rep. Washington.

Petrović, Saša, Miles Osborne, and Victor Lavrenko (2010). "Streaming first story detection with application to Twitter". In: *Proceedings of The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Stroudsburg: Association for Computational Linguistics, pp. 181–189.

Qu, Yan, Philip Fei Wu, and Xiaoqing Wang (2009). "Online Community Response to Major Disaster: A Study of Tianya Forum in the 2008 Sichuan Earthquake". In: *Proceedings of the 42nd Hawaii International Conference on System Sciences*. IEEE, pp. 1–11.

Quarantelli, E.L. (2008). "Disaster crisis management: A summary of research findings". In: *Crisis Management* 2, pp. 45–56.

Queensland Government (2013). *Disaster management phases*. http://www.disaster.qld.gov.au/About_Disaster_Management/Management_Phases.html. Accessed: 2014 Feb.

Raban, Daphne Ruth (2008). "The Incentive Structure in an Online Information Market". In: *Journal of the American Society for Information Science and Technology* 59.14.

Rogstadius, Jakob et al. (2011a). "A real-time social media aggregation tool: Reflections from five large-scale events". In: *Workshop on CSCWSmart at ECSCW*. Aarhus.

Rogstadius, Jakob et al. (2011b). "An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets". In: *Proc. AAAI Conference on Weblogs and Social Media*. Vol. 11. Barcelona.

Rogstadius, Jakob et al. (2011c). "Towards Real-time Emergency Response using Crowd Supported Analysis of Social Media". In: *CHI Workshop on Crowdsourcing and Human Computation*. Vancouver: ACM Press.

Rogstadius, Jakob et al. (2013a). "An Introduction for System Developers to Volunteer Roles in Crisis Response and Recovery". In: *In Proceedings of the International Conference on Information Systems for Crisis Response and Management*. Baden-Baden.

Rogstadius, Jakob et al. (2013b). "CrisisTracker: Crowdsourced social media curation for disaster awareness". In: *IBM Journal of Research and Development* 57.5, 4:1–4:13.

Salganik, Matthew J., Peter Sheridan Dodds, and Duncan J. Watts Watts (2006). "Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market". In: *Science* 311.5762, pp. 854–856.

SBTF (2012). *Introducing the Standby Task Force*. http://blog.standbytaskforce.com/about/introducing-the-standby-task-force/. Accessed: 2012 Oct.

SBTF & UN OCHA (2011). *Libya Crisis Map Deployment*. Tech. rep.

Schrader, Y. John (1993). *The Army's role in domestic disaster support: An assessment of policy choices*. Tech. rep. Santa Monica.

Stallings, A. and E.L. Quarantelli (1985). "Emergent citizen groups and emergency management". In: *Public Administration Review* 45, pp. 93–100.

Starbird, Kate and Leysia Palen (2010). "Pass It On?: Retweeting in Mass Emergency". In: *Proc. ISCRAM '10*. Seattle, USA.

Starbird, Kate et al. (2010). "Chatter on The Red; What Hazards Threat Reveals about the Social Life of Microblogged Information". In: *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. New York: ACM, pp. 241–250.

Stelter, Brian and Noam Cohen (2008). *Citizen Journalists Provided Glimpses of Mumbai Attacks*. http://www.nytimes.com/2008/11/30/world/asia/30twitter.html?_r=0. Accessed: 2013 Jan 8.

Sutton, Jeanette, Leysia Palen, and Irina Shklovski (2008). "Backchannels on the Front Lines: Emergent Uses of Social Media in the 2007 Southern California Wildfires". In: *Proceedings of the 5th International ISCRAM Conference*. Washington, D.C.

The Guardian (2011). *How riot rumours spread on Twitter*. http://www.theguardian.com/uk/interactive/2011/dec/07/london-riots-twitter. Accessed: 2012 Feb 5.

Torrey, Cristen et al. (2006). "Connected Giving: Ordinary People Coordinating Disaster Relief on the Internet". In: *Proceedings of the 40th Annual Hawaii International Conference on System Sciences*. Big Island, Hawaii: IEEE Computer Society, 179a.

Turrell, Chris (2011). *Social Media 101*. http://www.cmlor.com/blog/social-media-101/. Accessed: 2013 Jan 8.

Twitter, Inc. (2013). *Amendment no. 1 to form S-1*. http://www.sec.gov/Archives/edgar/data/1418091/000119312513400028/d564001ds1a.htm.

United Nations (2013). *What are UN Clusters?* http://business.un.org/en/documents/249. Accessed: 2013 Dec 2.

United Nations Volunteers (2011). *2011 State of the world's volunteerism report*. Tech. rep.

Vasterman, Peter, C. Joris Yzermans, and Anja J. E. Dirkzwager (2005). "The Role of the Media and Media Hypes in the Aftermath of Disasters". In: *Epidemiologic Reviews* 27.1, pp. 107–114.

Verity, Andrej (2011). *OCHA's Lessons Learned: Collaboration with V&TCs for Libya and Japan*. Tech. rep.

– (2013). *These are the Humanitarian Decision Makers*. http://blog.veritythink.com/post/60157407408/these-are-the-humanitarian-decision-makers. Accessed: 2013 Nov 29.

Vieweg, Sarah (2012). *Situational Awareness in Mass Emergency: A Behavioral and Linguistic Analysis of Microblogged Communications*. Tech. rep.

Vieweg, Sarah et al. (2008). "Collective Intelligence in Disaster: An Examination of the Phenomenon in the Aftermath of the 2007 Virginia Tech Shootings". In: *Proceedings of the 5th International ISCRAM Conference*. Washington, D.C.

Vieweg, Sarah et al. (2010). "Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness". In: *Proc. CHI 2010*, pp. 1079–1088.

Vukovic, Maja (2009). "Crowdsourcing for Enterprises". In: *IEEE Congress on Services*. Los Angeles, pp. 686–692.

Wiharta, Sharon et al. (2008). *The Effectiveness of Foreign Military Assets in Natural Disaster Response*. Tech. rep.

Yin, Jie et al. (2012). "Using Social Media to Enhance Emergency Situation Awareness". In: *IEEE Intelligent Systems* 27.6, pp. 52–59.

A Nossa Universidade

Colégio dos Jesuítas
Rua dos Ferreiros - 9000-082, Funchal

Tel: +351 291 209400
Fax: +351 291 209410
Email: gabinetedareitoria@uma.pt